Full notes introduction to statistics

Introduction to statistics (Chinhoyi University of Technology)

Studocu is not sponsored or endorsed by any college or university
Downloaded by Khann (moin24126@gmail.com)

SHOUT
SHARE
IT

Contents

1 Introduction **1**
- 1.1. Overview of Statistics . . . 1
- 1.2. Definition of terms . . . 1
- 1.3. Sampling Techniques . . . 4
- 1.4. Probability Sampling methods . . . 5
  - 1.4.1. Simple Random Sampling . . . 5
  - 1.4.2. Systematic Random Sampling . . . 6
  - 1.4.3. Stratified Sampling . . . 6
  - 1.4.4. Cluster Sampling . . . 7
- 1.5. Non-probability sampling methods . . . 8
  - 1.5.1. Convinience or Availability . . . 8
  - 1.5.2. Quota / Proportionate . . . 8
  - 1.5.3. Expert or Judgemental . . . 8
  - 1.5.4. Chain referral / Snowballing / Networking . . . 9
- 1.6. Errors in sampling . . . 9
- 1.7. Data Collection Methods . . . 10
  - 1.7.1. Observation . . . 10
  - 1.7.2. Interview . . . 10
  - 1.7.3. Experimentation . . . 11

2 Data and Data Presentation **1**
- 2.1. Introduction . . . 1
- 2.2. Data Types . . . 1
  - 2.2.1. Qualitative random variables . . . 1
  - 2.2.2. Quantitative random variables . . . 2
- 2.3. Data sources . . . 6
- 2.4. Data presentation . . . 8
  - 2.4.1. Pie Charts . . . 8
  - 2.4.2. Bar Chart . . . 9
  - 2.4.3. Histograms . . . 9
  - 2.4.4. Stem and leaf diagram . . . 10
  - 2.4.5. Frequency Polygons . . . 11
- 2.5. Exercises . . . 11

3 Measures of Central Tendency **13**
- 3.1. Introduction . . . 13
- 3.2. Measures of Central Tendency . . . 13
- 3.3. Arithmetic Mean . . . 13
  - 3.3.1. Mean for ungrouped data . . . 14

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

3.3.2. Mean for grouped data . . . 14
3.4. The Mode . . . 15
3.4.1. Mode for ungrouped data . . . 15
3.4.2. Mode for grouped data . . . 16
3.5. The Median . . . 16
3.5.1. Median for ungrouped data . . . 17
3.5.2. Median for grouped data . . . 17
3.6. Quartiles . . . 18
3.6.1. Quartiles for ungrouped data . . . 19
3.6.2. Quartiles for grouped data . . . 19
3.6.3. The second quartile, Q_2 (Median) . . . 20
3.6.4. The upper quartile, Q_3 . . . 20
3.6.5. Percentiles . . . 21
3.7. Skewness . . . 21
3.8. Kurtosis . . . 22
3.9. Exercises . . . 22

4 Measures of Dispersion **23**
4.1. Introduction . . . 23
4.2. Range . . . 23
4.3. Variance . . . 24
4.4. Standard deviation . . . 27
4.5. Coefficient of variation . . . 27
4.6. Exercises . . . 28

5 Basic Probability **31**
5.1. Introduction . . . 31
5.2. Definition . . . 31
5.3. Approches to probability theory . . . 31
5.4. Properties of probability . . . 33
5.5. Basic probability concepts . . . 34
5.6. Types of events . . . 34
5.7. Laws of probability . . . 35
5.8. Types of probabilities . . . 37
5.9. Contigency Tables . . . 39
5.10. Tree diagram . . . 40
5.11. Counting rules . . . 40
5.11.1. Multiplication Rule . . . 40
5.11.2. Permutations . . . 41
5.11.3. Combinations . . . 41
5.12. Exercise . . . 42

6 Probability Distributions **45**
6.1. Introduction . . . 45
6.2. Definition . . . 45
6.3. Random variables . . . 45
6.4. Discrete probability distribution . . . 46
6.5. Properties of discrete probability mass function . . . 47
6.6. Probability terminology and notation . . . 48

Downloaded by Khann (moin24126@gmail.com)

6.7. Discrete probability distributions 49
6.7.1. Bernoulli distribution 49
6.7.2. Binomial distribution 50
6.7.3. Poisson distribution 51
6.8. Continuous probability distributions 53
6.8.1. The Normal distribution 53
6.8.2. The standard normal distribution 54
6.8.3. The Uniform distribution 56
6.8.4. The Exponential distribution 57

7 Interval Estimation **59**
7.1. Introduction 59
7.2. Confidence Intervals 59
7.3. Confidence Interval for the Population Mean 60
7.4. One-Sided Confidence Intervals for the Population Mean 63
7.5. Confidence Interval for the Population Proportion 67
7.6. Confidence Interval for the Population Variance 68
7.7. Confidence Interval for the Population Standard Deviation 69
7.8. Confidence Interval for the Difference of Two Populations Means 70
7.8.1. Case 1: Known Population Variance 70
7.8.2. Case 2: Unknown (but assumed Equal) Population Variances . . 70

8 Hypothesis Testing **73**
8.1. Important Definitions, and Critical Clarifications 73
8.2. General Procedure on Hypotheses Testing 75
8.3. Hypothesis Testing Concerning the Population Mean 75
8.3.1. Case 1: Known Population Variance 75
8.3.2. Guidelines to the Expected Solution 76
8.3.3. Case 2: Unknown Population Variance 76
8.4. Hypothesis Testing concerning the Population Proportion 78
8.5. Comparison of Two Populations 79
8.5.1. Hypothesis Testing concerning the Difference of Two Population Means 79
8.6. Independent Samples and Dependent/ Paired Samples 80
8.6.1. Advantages of Paired Comparisons 81
8.6.2. Disadvantages of Paired Comparisons 82
8.7. Test Procedure concerning the Difference of two Population Proportions 82
8.8. Tests for Independence 84
8.9. Ending Remark(s) 85

9 Regression Analysis **87**
9.1. Introduction 87
9.2. Uses of Regression Analysis 87
9.3. Abuses of Regression Analysis 88
9.4. The Simple Linear Regression Model 89
9.4.1. The Scatter Plot 89
9.4.2. The Correlation Coefficient 89
9.4.3. Regression Equation 90
9.4.4. Coefficient of Determination, r^2 91

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

10 Index numbers **95**
10.1. Objectives 95
10.2. Introduction 95
10.3. What is an Index Number? 95
10.3.1. Characteristics of an Index Numbers 95
10.3.2. Uses of Index Numbers 96
10.4. Types of Index Numbers 97
10.5. Methods of constructing index numbers 98
10.5.1. Aggregate Method 98
10.5.2. Merits and demerits of this method 99
10.5.3. Weighted Aggregates Index 100
10.5.4. Laspeyres Method 100
10.5.5. Merits and demerits of Laspeyres method? 101
10.5.6. Paasches Method 102
10.5.7. Merits and Demerits of Paasches Index 102
10.6. Fisher Index 104

Downloaded by Khann (moin24126@gmail.com)

Chapter 1

Introduction

1.1. Overview of Statistics

Statistics is when individual data values are collected, summarized,analysed and presented and used for decision making. It is an important tool in transforming raw data into meaning and usable information. Also statistics can be regarded as a decision support tool. A table below shows a transformation process of data to information.

Input	**Process**	**Output**
Data	*Statistical Analysis*	*Information*
Raw observation	Transformation process	Useful, Usable and Meaningful

An understanding of statistics allows managers to: i) Perform simple statistical analysis. ii) Intelligently prepare and interpret reports expressed in numerical terms. iii) Communicate effectively with statistical analysts. iv) Good decision making.

1.2. Definition of terms

The following terms shall be used in this module more often.

Statistics
Definition 1
Statistics refers to the methodology [collection techniques] for collection, presentation and analysis of data and the use of such data [Neter J. et al (1988)].

Definition 2
In common usage, it refers to numerical data. This means an y collection of data or information constitutes what is referred to as Statistics. Some examples under this definition are:

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

1. Vital statistics - These are numerical data on births, marriages, divorces, communicable diseases, harvests, accidents etc.

2. Business and economic statistics - These are numerical data on employment, production, prices, sales, dismissals etc.

3. Social statistics - These are numeric data on housing, crime, education etc.

Definition 3 - Statistics is making sense of data.

In Statistics (as in real life), we usually deal with large volumes of data making it difficult to study each observation (each data point), in order to draw conclusions about the source of the data. We seek a statistical method or methods that can summarise the data so that we can draw conclusions about these data, without scrutinising each observations (which is rather difficult). Such methods fall under area of statistics called **descriptive** statistics.

A **Statistician** is an individual who collects dat, analyses it using statistical techniques, interprets the results and makes conclusions and recommendations on the basis of the data analysis.

Parameter(s) - These are numeric measure(s) derived from a population e.g. population mean (μ), population variances (σ^2), and population standard deviation (σ).

Data

Data is what is more readily available from a variety of sources and of varying quality and quantity. Precisely data is individual observations on an issue and in itself conveys no useful information.

Information

To make sound decision, one needs good and quality information. Information must be timely, accurate, relevant, adequate and readily available. Information is defined to processed data. Table above summarizes relationship between data and information. GIGO - Garbage In Garbage Out.

Random variable

A variable is any characteristic being measured or observed. Since a variable can take on different values at each measurement it is termed a random variable. For example, Sales, Company turnover, Weight, Height, yield, Number of babies born, e.t.c

Downloaded by Khann (moin24126@gmail.com)

Population
A population is a collection of elements about which we wish to make an inference. The population must be clearly defined before the sample is taken.

Target population
The population whose properties are estimated via a sample or usually the 'total' population.

Sample
A sample is a collection of sampling units drawn from a population. Data are obtained from the sample and are used to describe characteristics of the population. A sample can also be defined as a subset / part of or a fraction of a population.

statistic(s)
These are numeric measure(s) derived from a sample e.g. sample mean ($\bar{x}$), sample variances (s^2), and sample standard deviation (s).

Sampling Frame
A sampling frame is a list of sampling units. A set of information used to identify a sample population for statistical treatment. It includes a numerical identifier for each individual, plus other identifying information about characteristics of the individuals, to aid in analysis and allow for division into further frames for more in-depth analysis.

Sampling.

A process used in statistical analysis in which a predetermined number of observations will be taken from a larger population. The methodology used to sample from a larger population depend on the type of analysis being performed, but includes simple random sampling, systematic sampling and observational sampling. These will be discussed later.

Sampling Units
Sampling units are non-overlapping collections of elements from the population that cover the entire population. It is a member of both the sampling frame and of the sample. The sampling units partition the population of interest for example households or individual persons for census.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

1.3. Sampling Techniques

We do explore the sampling techniques in order to be able to decide which one is the most appropriate for each given situation. Sampling techniques are methods of how data can be collected from the given population.

Types of Sampling

Probability Sampling
Has a distinguishing characteristic that each unit in the population has a known, nonzero probability of being included in the sample thus, it is clear that every subject or unit has an equal chance of being selected from the population. These probabilities are usually equal. It eliminates the danger of being biased in the selection process due to one's own opinions or desires.

Non-probability Sampling
Is a process where probabilities cannot be assigned to the units objectively, and hence it becomes difficult to determine the reliability of the sample results in terms of probability. A sample is selected according to one's convenience, or generality in nature. It is a good technique for pilot or feasibility studies. Examples include purposive sampling, convenience sampling, and quota sampling. In non-probability sampling, the units that make up the sample are collected with no specific probability structure in mind e.g. units making up the sample through volunteering.

Remark:
We shall focus on probability sampling because if an appropriate technique is chosen, then it assures sample representativeness and hence the errors for the sampling can be estimated.

Reasons to use Sampling
Sampling is done mostly for reasons of Cost, Time, Accessibility, Utility and Speed. Expansion on the reasons is left for the lecture. Some points to clearly define when sampling. Sampling method to be employed.

Sample size
Reliability degree of the conclusions that we can obtain i.e. an estimation of the error that we are going to have. An inappropriate selection of the elements of the sample

Downloaded by Khann (moin24126@gmail.com)

can cause further errors once we want to estimate the corresponding population parameters.

1.4. Probability Sampling methods

The four methods of probability sampling are simple random, systematic, stratified and cluster sampling methods.

1.4.1. Simple Random Sampling

Requires that each element of the population have an equal chance of being selected. A simple random sample is selected by assigning a number to each element in the population list and then using a random number table to draw out the elements of the sample. The element with the number drawn out makes it into the sample. The population is ”mixed up” before a previously specified number, n, of elements is selected at random. Each member of the population is selected one at a time, independent of one another. However, it is noted that all elements of the study population are either physically present or listed.
Also, regardless of the process used for this method, the process can be laborious especially when the list of the population is long or it is completed manually without the aid of a computer. A simple random sample can be got using calculator (random key), computer (using excel function =rand()), or random number tables.

In this method, every set of n elements in the population has an equal chance of being selected as the sample.

Advantages

- It eliminates bias due to the personal judgment or discretion of the researcher
- More representative of the population
- Estimates are more accurate

Demerits

- Requires an up to date sampling frame
- Numbering of the population elements may be time consuming eg for large populations

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Illustration
An example of simple random sampling may include writing each member of the population on a piece of paper and putting in a hat. Selecting the sample from the hat is random and each member of the population has an equal chance of being selected. However, this approach is not feasible for large populations, but can be completed easily if the population is very small.

1.4.2. Systematic Random Sampling

Selection of sampling units is done in sequences separated on lists by the interval selection. In this method, every n^{th} element from the list is selected as the sample, starting with a sample element n randomly selected from the first k elements. For example, if the population has 1000 elements and a sample size of 100 is needed, then k would be $\frac{1000}{100} = 10$. Now, if the number 7 is randomly selected from the first ten elements on the list, then the sample would continue down the list selecting the 7^{th} element from subsequent groups of ten elements. Care must be taken when using systematic sampling to ensure that the original population list has not been ordered in a way that introduces any non-random factors into the sampling.

Illustration
An example of systematic sampling would be if an official from the Academic Registry of a hypothetical university is to register students for a tour of regional universities. The official may select at random the 15^{th} student out of the first 20 students in a list of all students in the university. This official would then keep adding twenty and selecting the 35^{th} student, 55^{th} student, 75^{th} student and so on to register for the tour of regional universities until the end of the list is reached.

Remark:
In cases where the population is large and the population list is available, systematic sampling is usually preferred over simple random sampling since it is more convenient to the experimenter.

1.4.3. Stratified Sampling

It is used when representatives from each homogeneous subgroup within the population need to be represented in the sample. The first step in stratified sampling is to divide the population into subgroups (strata) based on mutually exclusive criteria. Random or systematic samples are then taken from each subgroup. The *sampling fraction* for each subgroup may be taken in the same proportion as the subgroup has in the population.

Downloaded by Khann (moin24126@gmail.com)

Illustration
As an example, if an owner of a local supermarket conducting a customer satisfaction survey may wish to select random customers from each customer type in proportion to the number of customers of that type in the population. Suppose 40 sample units are to be selected, and 10% of the customers are managers, 60% are users, 25% are operators and 5% are students from CUT, then 4 managers, 24 users, 10 operators and 2 students from CUT would be randomly selected.

Remark:
Stratified sampling can also sample an equal number of items from each subgroup.

1.4.4. Cluster Sampling

In cluster sampling, the population that is being sampled is divided into naturally occurring groups called clusters. A cluster is as heterogeneous as possible to matching the population clusters which says that a cluster is representative of the population. A random sample is then taken from within one or more selected.

Illustration
An organization with 300 small branches providing a service country wide has an employee at the HQ who is interested in auditing for compliance to some coding standard. The employee might use cluster sampling to randomly select 40 branches as representatives for the audit and then randomly sample coding systems for auditing from just the 40.

Remark:
Cluster sampling can tell us a lot about that particular cluster, but unless the clusters are selected randomly and a lot of clusters are sampled, generalizations cannot always be made about the entire population.

Difference between a Cluster & a Stratum
A cluster is a heterogeneous subgroups but a stratum is a homogeneous subgroup. A summary of probability sampling methods is discussed below.

Simple random

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Each member of the study population has an equal probability of being selected.

Systematic
Each member of the study population is either assembled or listed, a random start is designated, then members of the population are selected at equal intervals

Stratified
Each member of the study population is assigned to a homogeneous subgroup or stratum, and then a random sample is selected from each stratum.

Cluster
Each member of the study population is assigned to a heterogeneous subgroup or cluster, then clusters are selected at random and all members of a selected cluster are included in the sample.

1.5. Non-probability sampling methods

There are four methods of non-probability sampling discussed in this module that are Convienience, Quota, expert and chain referral.

1.5.1. Convinience or Availability

This is sampling which is based on the proximity of the population elements to the decision maker. Being at the right place at the right time. Elements nearby are selected, and those not in close physical or communication range are not considered.

1.5.2. Quota / Proportionate

This is sampling in which certain distinct or known characteristics in the population should appear in relatively similar proportions. Eg A population (N) of 100 comprised of 60 being females and 40 being males. If a sample of n=20 is to be selected then that ratio of 6:4 has to be reflected.

1.5.3. Expert or Judgemental

This is sampling in which the decision maker has direct or indirect control over which elements are to be included in the sample. Appropriate when the decision maker feels

Downloaded by Khann (moin24126@gmail.com)

that some population members have better or more information than others. Or some members are more representative than others.

1.5.4. Chain referral / Snowballing / Networking

The researcher starts with a person who displays qualities of interest then refers to the next and so on.

1.6. Errors in sampling

During sampling, errors can be committed by the statistician. These are either **sampling** or **non-sampling** errors. Errors can be corrected by sampling without bias. Some common sources of bias are i) incorrect sampling operation, non-interviews. Some errors that arises in sampling are discussed below.

Selection error
Selection error occurs when some elements of the population have a higher probability of being selected than others. Consider a scenario where a manager of a local supermarket wishes to measure how satisfied his customers. He proceeds to interview some of them from 08:00 to 12:00. Clearly, the customers who do their shopping in the afternoon are left out and will not be represented making the sample unrepresentative of all the customers. Such kind of errors can be avoided by choosing the sample so that all the customers have the same probability of being selected. This is a sampling error.

Non-Response Error
It is possible that some of the elements of the population do not want or cannot answer certain questions. It may also happen when we have a questionnaire including personal questions, that some of the members of the population do not answer honestly or would rather avoid answering. These errors are generally very complicated to avoid, but in case that we want to check honesty in answers, we can include some questions called filter questions to detect if the answers are honest. This is a non-sampling error.

Interviewer influence
The interviewer may fail to be impartial i.e. s/he can promote some answers more than others.

Remark:
A sample that is not representative of the population is called a biased sample.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Questions relating to selecting out of naturally arise. These are: *When concluding about the population, how many of the population elements is represented by each one of the sample elements?* What proportion of the population are we selecting? The responses lie in the following factors.

1.7. Data Collection Methods

The three data collection methods are: **Observation,Interview and Experimentation.** Depending on the type of research and data to be collected, different methods can be used to collect that data set.

1.7.1. Observation

This method has the *direct* and *desk research* methods. **Direct observation** involves collecting data by observing the item in action. Examples for this method are: pedestrian flow, vehicle traffic, purchase behavior of a commodity in a shop, quality control inspection e.t.c. An **advantage** of this method is that the respondent behaves in a natural way since he is not aware that he is being observed. A **disadvantage** is that it is a passive form of data collection. Also there is no opportunity to investigate the behavior further. Desk research involves consulting and extracting secondary data from source documents and collect data from them.

1.7.2. Interview

This method collects primary data through direct questioning. A questionnaire is the instrument used to structure the data collection process. Three approaches in data collection using interviews are: *personal, postal and telephone interviews.*

Personal Interviews
A questionnaire is completed through face-to-face contact with the respondent. **Advantages** for this method are: High response rate, it allows probing for reasons, data collection is immediate, data accuracy is assured, useful for technical data is required, non-verbal responses can be observed and noted, more questions can be asked, responses are spontaneous, and use of aided-recall questions is possible. **Disadvantage** of this method are that it is time consuming, it requires trained interviewers, fewer interviews are conducted because of cost and time constraints, biased data can be collected if interviewer is inexperienced.

Downloaded by Khann (moin24126@gmail.com)

Postal Surveys

When target population is large and or geographically dispersed then use of postal questionnaires is considered most suitable. **Advantages** of this method is that larger sample of respondents can be reached, more costs effective, interviewer bias is eliminated, respondents have more time to consider their responses, anonymity of respondents is assured resulting in more honest responses, respondents are more willing to answer personal questions. The **disadvantages** for this method are: low response rate, respondents cannot get clarity on some questions, mailed questionnaires must be short and simple to complete, limited possibilities of probing or further investigations, data collection takes long time, no control of who answers the questionnaire, and no possibilities of validating responses.

Telephone Interviews

The interview is conducted telephonically with the respondent. **Advantages** of this method are: it allows quicker contact with geographically dispersed respondents, callbacks can be made if respondent is not initially available, low cost, interviewer probing is possible, clarity on questions can be provided by the interviewer, and a larger sample of respondents can be reached in short space of time. **Disadvantages** are that respondent anonymity is lost, non-verbal responses cannot be observed, trained interviewers are required hence more costly, possible interviewer bias, respondent may terminate interview prematurely, and sampling errors are compounded if more respondents do not have telephones.

1.7.3. Experimentation

This is when primary data is generated through manipulation of variables under controlled conditions. The method is mostly used in scientific and engineering research. Data on the primary variable under study is monitored and recorded whilst the researcher controls effects of a number of influencing factors. Examples include: Demand elasticity for a product, advertising effectiveness. **Advantages** of this method are: quality data is collected and results are generally more objective and valid. The **disadvantages** are that the method is costly and time consuming, and may be impossible to control for certain factors which affects the results.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 2

Data and Data Presentation

2.1. Introduction

A Statistician collects data (in an appropriate manner) analyses it using statistical techniques, interprets the results and makes conclusions and recommendations on the basis of data analysis. The word *data* keeps turning in our discussion. Data is the *"blood of statistics"*.

The world of statistics resolves around data, there is no statistics without data. What is data? How is it collected? Why do we collect it? These are the questions to be answered in this chapter.

2.2. Data Types

An understanding of nature of data is necessary for 2 reasons. It enables a user: assess data quality and to select the appropriate statistical method to use to analyse the data.

Quality of data is influenced by three factors that are: **type, source** and **method** used to collect data. The type of data gathered determines the type of analysis which can be performed on the data. Certain statistical methods are valid for certain data types only. An incorrect application of a statistical method to a particular data type can render the findings invalid.

Data type is determined by the nature of the random variables which the data represents. Random variables are essentially of two kinds that are *Qualitative and Quantitative*.

2.2.1. Qualitative random variables

These are variables which yield categorical (non-numeric) responses. The data generated by qualitative random variables are classified into one of a number of categories.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

The numbers representing the categories are arbitrary i.e. codes: Coded values cannot be manipulated arithmetically, as it does not make sense.

Examples of qualitative random variables

Random variables	Response Categories	Data Codes
Managerial Level	Supervisor	1
	Section Head	2
	Departmental Head	3
	General Manager	4
Do you like soft drink?	Yes	2
	No	1
Gender	Female	0
	Male	1

2.2.2. Quantitative random variables

Quantitative random variables are variables that yield numeric responses. The data generated for quantitative random variables can be meaningfully manipulated using conventional arithmetic operations.

Examples of quantitative random variables

Random Variables	Response Range	Data
Age of employee	17 - 65 years	e.g. 39 years
Distance to work	0 - 20 km	e.g. 5.3 km
Class size	1, 2, 3 ...	e.g. 15 pupils

Each random variable category is associated with a different type of data. There are two classifications of data types.

Data type 1 - Data measurement scales

Data measurement scales include Nominal, Ordinal, Interval and Ratio-scaled data.

Nominal-scaled data

Objects or events are distinguished on the basis of a name. Nominal-scaled data is associated mainly with **qualitative** random variables. Where data of qualitative random variables is assigned to one of a number of categories of equal importance, then such data is referred to as nominal-scaled data. There is no implied ordering between the groups of the random variable.

Downloaded by Khann (moin24126@gmail.com)

Examples of nominal-scaled data
Table below shows examples of nominal scaled data.

Qualitative Random Variables	Response Categories	Data Code
Gender	Male / Female	1 / 2
Car type owned	Mazda/Golf/Toyota/Honda	1 / 2 / 3 / 4
City leaved in	Harare/Byo/Mutare/Gweru	1 / 2 / 3 / 4
Marital Status	Married/Single/Divorced/Widow	1 / 2 / 3 / 4
Engineering Profession	Civil/Electrical/Mechanical	1 / 2 / 3

Each observation of the random variables is assigned to only one of the categories provided. Arithmetic calculations cannot be meaningfully performed on the coded values assigned to each category. They are only numeric codes which are arbitrarily assigned and can be counted. Nominal-scaled data is the weakest form of data, since only a limited range of statistical analysis can be formed on such data.

Ordinal-scaled data
Objects or events are distinguished on the basis of the relative amounts of some characteristics they posses. The magnitude between measurements is not reflected in the rank. Such data is associated mainly with qualitative random variables. Like nominal-scaled data, ordinal-scaled data is also assigned to only one of a number of coded categories, but there is now a ranking implied between the categories in terms of being better, bigger, longer, older, taller, or stronger, etc. While there is an implied difference between the categories, this difference cannot be measured exactly. That is, the distance between categories cannot be quantified nor assumed to be equal. Ordinal-scaled data is generated from ranked responses in market research studies.

Examples of Ordinal-scaled data

Qualitative Random Variables	Response Categories	Data Codes
T-Shirt size	Small / Medium / Large	1 / 2 / 3
Company turnover	Small / Medium / Large	1 / 2 / 3
Management levels	Lower / Middle / Senior	1 / 2 / 3
Work experience	Little / Moderate / Extensive	1 / 2 / 3
Magazine type	Rank the top three magazine you often read	1 / 2 / 3
Sizes of bulbs	Smallest / Small / Large / Largest	1 / 2 / 3 / 4

There is a wider range of valid statistical methods (i.e. the area of non-parametric statistics) available for the analysis of ordinal-scaled data than there is for nominal-scaled data. Ordinal-scaled data is also generated from a ”counting process”.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Interval-scaled data

Interval-scaled data is associated with **quantitative** random variables. Differences can be measured between values of a quantitative random variable. Thus interval-scaled data possesses both order and distance properties. Interval-scaled data, however, does not possess an absolute origin. Therefore the ratio of values cannot be meaningfully compared for interval-scaled data. The absolute difference makes sense when interval-scaled data has been collected.

Examples of Interval-scaled data

Suppose four places A, B, C and D have temperatures $20^oC, 25^oC, 35^oC$ and 40^oC respectively. Using interval scale we see that the difference between A and B is equal to that of C and D. However ratios are not used. A value of 0^oC does not mean absence of temperature, also it is not correct to say temperature of D is twice as much as that of A.

Interval-scaled data is most often generated in marketing studies through rating responses on a continuum scale. A wide range of statistical techniques can be applied to interval scaled data as it posses numeric (measurement) properties.

Ratio-scaled data

This data is associated mainly with *quantitative* random variables. If the full range of arithmetic operations can be meaningfully performed on the observations of a random variable, the data associated with that random variable is termed ratio-scaled. It is a numeric data with a zero origin. The zero origin indicates the absence of the attribute being measured.

Example 1 of Ratio-scaled data

Quantitative Random Variable	Example of data values
Age	42 years
Income	$2,500
Distance	35 km
Time	32 minutes
Mass	240g
Price	$7.82

Such data are the strongest form of statistical data which can be gathered and lends itself to the widest range of statistical methods. Ratio-scaled data can be manipulated meaningfully through normal arithmetic operations. Ratio-scaled data is

Downloaded by Khann (moin24126@gmail.com)

gathered through a measurement process. It should be noted that if ratio-scaled data is grouped into categories, the data type becomes ordinal-scaled. This then reduces the scope for statistical analysis on the random variable.

Example 2 of Ratio-scaled data
Note: By capturing *Age* data in categories instead of actual age, the data becomes ordinal-scaled. However, the random variable remains quantitative in nature. See table below.

Random Variable	Response Category	Data code used
Age	0 - 16	1
	17 - 24	2
	25 - 36	3
	37 - 45	4
	46 - 55	5

When data capturing instruments are set up, care must be exercised to ensure that the most useful form of data is captured. However, this is not always possible for reasons of convenience, cost and sensitivity of information. This applies particularly to random variables such as age, personal income, company turnover and consumer behavior questions of a personal nature. The functional area of marketing generates mostly categorical (i.e. nominal/ordinal) data arising from consumer studies, while the areas of finance/accounting and production generate mainly quantitative (ratio) data. Human resources management generates a mix of qualitative and quantitative data for analysis.

Data type 2

A second classification of data type is either **discrete** and **continuous** data.

Discrete data
A random variable whose observations can take on only specific values, usually only integer values, is referred to as a discrete random variable. In such instances, certain values are valid, while others are invalid.

Examples of random variables generating discrete data
(i) Number of cars in a parking lot at a given time, (ii) Daily number of hotel rooms booked for January 1992, (iii) Number of students in a class, (iv) Number of employees in an organization, (v) Number of paintings in an art collection, (vi) Number of cars sold in a month by a dealer, (vii) Number of life assurance policies issued in 1990 in Zimbabwe.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Continuous data
A random variable whose observations take on any value in an interval is said to generate continuous data. This means that any value between a lower and an upper limit is valid.

Examples of random variables generating continuous data
(i) Time taken to travel to work daily, (ii) Age of a bottle of red wine, (iii) Mass of a caravan, (iv) Tensile strength of material, (v) Speed of an aircraft, (vi) Length of a ladder.

2.3. Data sources

Data for statistical analysis are available from any different sources. There are two classification types of data sources that are: Internal/external and Primary/secondary sources.

Internal data sources
This refers to the availability of data from within an organisation; internal data are generated during the course of normal business activities. Examples include: i) *Financial data* sales vouchers, credit notes, accounts receivable, accounts payable, asset register. ii) *Production data* - production cost records, stock sheets. iii) *Human Resource data* - time sheets, wages and salaries schedule, employee personal employment files. iv) *Marketing data* sales data, advertising expenditure.

External data sources
Data available from outside an organization is referred to as external data sources. Such sources may be private institutions, trade/employer/employee associations, profit motivated organizations or government bodies. The cost of the external data is dependent on the source. Generally, the cost is greater from private bodies than it is from government or public sources. Examples include: i) *Private source include* - Commercial and Industrial Association of Business, Research Bureaux. ii) *Public domain sources* include Newspapers, journals, trade magazines, reference material in libraries, The Central Statistical Services (ZimStats) is the Governments data capturing and dissemination instrument and others such as Universities, Reference Libraries, Banks' economic reports.

Primary data sources

Downloaded by Khann (moin24126@gmail.com)

Data which is captured at the point where it is generated is called Primary data. Such data is captured for the first time and with specific purpose in mind. Examples of data sources are: Largely the same as for internal data source, but also includes survey data (personnel surveys, salary surveys, market research surveys).

Advantages of primary data
Primary data are directly relevant to the problem at hand and generally offer greater control over data accuracy.

Disadvantages of primary data
Primary data can be time consuming to collect and are generally more expensive to collect (e.g. Market Research)

Secondary data sources
Data collected and processed by others for a purpose other than the problem at hand are called secondary data. Such data are already in existence either within or outside an organisation, i.e. one can get both internal secondary and external secondary data. The problem at hand determines whether the data are primary or secondary. Examples of internal secondary data are: Aged market research figures, previous financial statements of your company and past sales reports. Examples of external secondary data Reports produced by external data sources.

Advantages of secondary data
Some of the advantages of use of secondary data are: The data are already in existence, Access time is relatively short, The data are generally less expensive to acquire.

Disadvantages of secondary data
Some disadvantages of secondary data are:

- Data may not be problem specific.
- Data may be outdated and hence inappropriate.
- It may be difficult to assess data accuracy.
- Data may not be subject to further manipulation.
- Combining various sources could lead to errors of collation and introduce bias.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

2.4. Data presentation

Data can be presented in tables or graphs. Graphical techniques are pictorial or graphical representations of data such that the main features of the data are captured. The various graphical techniques which we will cover in this unit. Pie charts, bar charts, histograms, box and whisker plots and stem and leaf displays. Some other techniques which are important are dotplots, Lorenz curve and Z curves are not discussed in this module.

2.4.1. Pie Charts

A pie chart as the name suggests, is a circle divided into segments like a pie cut into pieces from the centre outwards. Each segment represents one or more values taken by a variable. Such charts are used to display qualitative data. Let us now look at an example, and see how we can construct and interpret a pie chart.

Example 1.1
The ages of 10 students doing BSCAC program at Chinhoyi University of Technology are: 26, 28, 28, 16, 22, 35, 42, 19, 55, 28. Grouping the ages into classes of 25 and below, 26-35, 36-45, and above 45, leads to a frequency distribution table below.

Age group	Number of Students
Below 25	3
26 - 35	5
36 - 45	1
Above 45	1

We now express these age groups as proportions or percentages and then indicate the angle in degrees as in table below.

Age group	Number of Students	Proportions	Percentages	Angle
Below 25	3	$\frac{3}{0+3+5+1+1} = \frac{3}{10}$	$\frac{3}{10} \times 100 = 30\%$	108^o
26 - 35	5	$\frac{5}{10}$	$\frac{5}{10} \times 100 = 50\%$	180^o
36 - 45	1	$\frac{1}{10}$	$\frac{1}{10} \times 100 = 10\%$	36^o
Above 45	1	$\frac{1}{10}$	$\frac{1}{10} \times 100 = 10\%$	36^o

There are only 4 groups. What we wish to do is to represent these percentages as angles in degrees i.e. instead of everything in Table 1.2 (column 2) above adding up to 10 (or 100 in the case of percentages) we want them to add up to 360^o (the total number of degrees in a circle) as shown in column 5. The calculation of the angle of

Downloaded by Khann (moin24126@gmail.com)

the i^{th} category can be done directly from the observations by using.

$$\frac{X_i}{\sum_{i=1}^{n} X_i} = 360^o$$

i.e. each observation multiplied by 3600 divided by the sum of the observations

2.4.2. Bar Chart

A bar chart, as the name suggests, is a visual presentation of data by means of bars or blocks put side by side. Each bar represents a count of the different categories of the data. Although both pie chart and bar graphs (as they are sometimes called) are used to illustrate qualitative data or discrete qualitative data, bar charts use the actual counts or frequencies of occurrences of each category of data. We need not use the actual data; we can use the percentage to come up with the Bar graph. Let us use the data in example 1.1 to illustrate the bar chart.

Example 1.2
We will now construct the bar chart using the data in example 1.1. We come up with suitable scales for the height and width of the graph, which are such that the graph is clear and representative in class example. The bars represent each age group count in terms of height. You can choose to make the bars thin or wide, it's up to you all you need to be certain of is that the bars represent each age group in terms of height. The bars should be of the sane width. Often, we represent each category by different colours or shades. This is especially useful when we are comparing several groups. For instance, we could be comparing the age groups of different intakes that would mean several graphs all put side by side. In this way we can compare the intakes aged X over different years.

2.4.3. Histograms

A histogram is a gragh drawn from a frequency distribution. It is used to represent continuous quantitative data. It usually consists of adjacent, touching rectangles or bars. The area of each rectangle is drawn in proportion to the frequency corresponding to that frequency class. When the class intervals are equal, the area of each rectangle is a constant multiple of height and so the histogram can be drawn as for a bar chart, except that the rectangles are touching. If the class intervals are not equal, the frequencies are adjusted accordingly to come up with frequency densities for the larger class intervals.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Exercise
Consider results of a test written by 45 students and marked out of 70. Data is presented in categories in table below. Use the data above to draw a histogram for the

Marks	Frequencies
10 - 19	7
20 - 29	20
30 - 39	9
40 - 49	3
50 - 59	5
60 - 69	1

mark distribution.

2.4.4. Stem and leaf diagram

A stem and leaf diagram is basically a histogram where the rectangles are built up to the correct height by individual numbers. Each data value is split up into its stem, the first digit [or first two digits, etc., depending on the data], and its leaves. Thus, the number 23 has stem 2 and leaf 3. The number 7 has stem 0 and leaf 7. Perhaps an example will illustrate this diagram.

Example 1.3
A scientist interested in finding out the age groups of people interested in cultural movies went to a movie theatre and collected the following information. Ages of people watching movie is shown below.

7	15	22	38	12	18	14	26	20	15	22	34	12	18	24
19	14	29	21	32	12	17	24	13	25	20	15	31	11	16
23	39	19	14	28	20	9	16	22	39	13	25	19	14	31

The stem 0, 1, 2 and 3 are listed on the left side of a vertical line and the leaves on the right side opposite the appropriate stem. The stem and leaf diagram of these data are represented below. A stem and leaf display should always have a **key** that indicates how data is displayed. Key: $0|7 = 7, 3|8 = 38$.

Table 2.1: Stem Plot of Ages, Key: $1|1 = 11$

Stem	Leaf
0	79
1	122233444455566788999
2	000122234455689
3	1124899

Also take note that 1^{st}, 2^{nd}, 3^{rd}, e.t.c. number on the leaf side should be in the same columns for the histogram feature to reveal.

Downloaded by Khann (moin24126@gmail.com)

2.4.5. Frequency Polygons

Frequency polygons are one alternative to histograms. The only difference here is that a frequency polygon is a line plot of the frequencies against the corresponding class mid-points. The points are joined by straight lines.

2.5. Exercises

1. Classify the following data sources as either Primary or secondary and Internal or external

   (a) The economic statistics quoted in The Financial Gazette.

   (b) The sum assured values on life assurance polices within your company.

   (c) The financial reports of all companies on the Zimbabwean stock exchange for the purpose of analyzing earnings per share.

   (d) Employment statistics published by ZimStats.

   (e) Market research findings on driving habits conducted by the ZRP Traffic section.

2. Define primary and secondary data. Include in your answers the advantages and disadvantages of both data types. Give two examples of secondary data.

3. What is the difference between primary and secondary data?

4. Areas of continents of the World

Continent	Area in millions of km^2
Africa	30.3
Asia	26.9
Europe	4.9
North America	24.3
Oceania	8.5
South America	17.9
Russia	29.5

   (a) Draw a bar chart of the above information.

   (b) Construct a pie chart to represent the total area.

5. The distance (km) travelled by a courier service motorcycle on 30 trips were recorded by the driver.

24	19	21	27	20	17	17	32	22	26	18	13	23	30	10
13	18	22	34	16	18	23	15	19	28	25	25	20	17	15

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

(a) Define the random variable, the data type and the measurement scale.

(b) From the data, prepare:

   i. an absolute frequency distribution

   ii. a relative frequency distribution and

   iii. the (relative) less than ogive.

(c) Construct the following graphs:

   i. a histogram of the relative frequency distribution,

   ii. stem and leaf diagram of the original data

(d) From the graphs, read off what percentage of trips were:

   i. between 25 and 30 km long

   ii. under 25km

   iii. 22km or more?

Downloaded by Khann (moin24126@gmail.com)

Chapter 3

Measures of Central Tendency

3.1. Introduction

From the previous unit, graphical displays were discussed. These are useful means of communicating broad overviews of the behaviour of a random variable. However, there is a need for numeral measures, called statistics, which will convey more precise information about the behaviour pattern of a random variable. The behaviour or pattern of any random variable can be described by measures of:

- Central tendency and
- Dispersion of observations about a central value.

3.2. Measures of Central Tendency

These are statistical measures which quantify where the majority of observations are concentrated. They are also called measures of location. A central tendency statistic represents a typical value or middle data point of a set of observations and are useful for comparing data sets. These measures may be based on the source that is whether they are from a population or from a sample. If from a population, we talk of a **parameter** and if from a sample we refer to **statistic**. The three main measures of central tendency are:

- Arithmetic mean or average
- Mode and
- Median

Each measure will be computed for ungrouped data and grouped data.

3.3. Arithmetic Mean

Given a set of n sample data values denoted by x_i for $i = 1, 2, 3, ...n$, the arithmetic mean is defined by $\bar{x}$. For a population, we refer to a population mean, μ

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

3.3.1. Mean for ungrouped data

The arithmetic mean for ungrouped data from a sample is defined as in:

$$\begin{aligned} \bar{x} &= \frac{sum\ of\ all\ observations}{total\ number\ of\ observations} \\ \bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned} \tag{3.1}$$

The population mean is defined as:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N}$$

Where n is the number of observations in the sample, x_i is the value of the i^{th} observation of random variable x and $\bar{x}$ is the symbol for a sample arithmetic mean. $\sum_{i=1}^{n} x_i$ is the shorthand notation for the sum of n individual observation i.e.
$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + ... + x_n$

3.3.2. Mean for grouped data

Grouped data is represented by a frequency distribution. All that is known is the frequency with which observations appear in each of the m classes. Thus the sum of all the observation cannot be determined exactly. Consequently, it is not possible to compute an exact arithmetic mean for the data set. The computed mean is an approximation of the actual arithmetic mean.

$$\bar{x} = \frac{\sum_{i=1}^{m} f_i x_i}{\sum f_i} \tag{3.2}$$

The population mean for grouped data is:

$$\mu = \frac{\sum_{i=1}^{m} f_i x_i}{N} \tag{3.3}$$

Where m is the number of classes in the frequency distribution. n is the number of observations in the sample x_i is the value of the i^{th} observation of random variable x, f_i is frequency of the i^{th} class, $\bar{x}$ is the symbol for a sample arithmetic mean. For you practice attempt examples in the Tutorial Work Sheet.

Downloaded by Khann (moin24126@gmail.com)

Properties of the Mean

- The arithmetic mean uses all values of the data set in its computation.
- The sum of deviation of each observation from the mean value is equal zero. i.e. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. This makes the mean an unbiased statistical measure of central location.

Drawbacks of the mean

- It is affected or distorted by extreme values (Outliers) in the data.
- It is not valid to compute the mean for *nominal- or ordinal-scaled* data. It is only meaningful to compute the arithmetic mean for *ratio-scaled* data (discrete or continuous)

There are other means that can be calculated for different distribution of values which are **Harmonic mean, Geometric mean** and **Weighted arithmetic mean**. We are not discuss them at this time.

3.4. The Mode

The *mode* of a given set of data is the observation with the *highest frequency*. In other words it is the *most occurring value* in a data set. It is that observation with the highest frequence. A distribution can have one mode (unimodal), two models (bimodal) or many modes(multimodal). Calculations for a mode from a population and sample is the same.

3.4.1. Mode for ungrouped data

In ungrouped data set, mode is obtained by observing the data carefully then finding the most frequently occurring observation. However, if the number of observations is too large, the mode can be found by arranging the data in ascending order and by inspection, identify that value that occurs frequently.

Example 1
If B=Blue, G=Green, R=Red and Y=Yellow. Consider a sample: YGBRBBRGYB, picked from a mixed bag. What is the modal colour?

Solution
The modal colour is Blue, because it appears most, with a frequency of 4.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

3.4.2. Mode for grouped data

In finding mode for grouped data, we first identify the modal class i.e. the class interval with the highest frequency. The mode lies in this class and then calculate the modal value using the formula

$$Mode = l_{mo} + \frac{c(f_1 - f_0)}{2f_1 - f_0 - f_2} \tag{3.4}$$

where l_{mo} is the lower limit of modal class, f_1 is the frequency of the modal class, f_0 is the frequency of the class preceeding the modal class, f_2 is the frequency of interval succeeding the modal class and c is the width of the modal class.

Example 2

Find the modal test mark for the following data.

Test mark, x	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30
Frequency	3	5	7	2	4

Solution
We seek to invoke the formula

$$Mode = l_{mo} + \frac{c(f_1 - f_0)}{2f_1 - f_0 - f_2}$$

where 15-20 is the modal class with the highest frequency of 7, $l_{mo} = 15$, $f_1 = 7$, $f_0 = 5$, $f_2 = 2$, and $c = 5$. Substituting these in the equation above yields

$$\begin{aligned} Mode &= 15 + \frac{5(7-5)}{2(7) - 5 - 2} \\ Mode &= 16.42 \end{aligned}$$

3.5. The Median

The *median* is the value of a random variable which divides an ordered (ascending or descending order) data set into two equal parts. It is also called the second quartile Q_2 or 50^{th} percentile. Half of the observation will fall below this median value and the other half above it. If the number of observations, n, is odd, then the median is observation $(\frac{n+1}{2})^{th}$. If the number of observations is even, then median is the mean of $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ observation. First, we consider a modified scenario of ungrouped data.

Downloaded by Khann (moin24126@gmail.com)

3.5.1. Median for ungrouped data

Given the following income data presented in a frequency distribution table, find the median.

Income ($)	3800	4100	4400	4900	5200	5500	6000
Number of workers	12	13	25	17	15	12	6

Solution
The number of observations is 100, which is even thus the median is mean of $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ observations i.e. the mean of the 50^{th} and 51^{st} observations. To find these observations we first find the cumulative frequencies of the data set. The 50^{th} observation is 4400 and 51^{st} observation is 4900. Thus

$$\begin{aligned} Median &= \frac{4400+4900}{2} \\ Median &= 4650 \end{aligned}$$

Interpretation
This means 50% of the workers get incomes that are less than \$4650 and another 50% get an income that is more than \$4650.

3.5.2. Median for grouped data

Given the following grouped data in a frequency table, find the median.

Income ($)	3601-3800	3801-4100	4101-4400	4401-4900	4901-5200	5201-5500	5501-6000
Number of workers	12	13	25	17	15	12	6
Cumulative frequency	12	25	50	67	82	94	100

We use a standard formular to calculate the *median* of the above grouped data which is

$$Median = O_{me} + \frac{c(\frac{n}{2} - F(<))}{f_{me}} \tag{3.5}$$

where m_e is the median class, O_{me} is the lower limit of median class, *n* is the sample size i.e. total number of observations, $F(<)$ is the cumulative frequency of class prior to median class and *c* is the width of the median class.

To use this formula, we calculate the cumulative frequencies and then identify the median class which is the class containing the $(\frac{n+1}{2})^{th}$ observation.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Example 4:
Calculate the median of following grouped data.

Mark	0-10	10-20	20-30	30-40	40-50
Frequency	2	12	22	8	6

Solution:
First and foremost, order the data set, in this case its already ordered. Then calculate the cumulative frequencies we get.

Mark	0-10	10-20	20-30	30-40	40-50
Frequency	2	12	22	8	6
Cumulative Frequency	2	14	36	44	50

We wish to use the formula

$$Median = O_{me} + \frac{c(\frac{n}{2} - F(<))}{f_{me}}$$

Where the median class is 20 - 30, c = 10, n = 50, $F(<) = 14$, f = 22, $f_{me} = 22$ and $O_{me} = 20$. Substituting we have

$$\begin{aligned} M_e &= 20 + \frac{10[\frac{50}{2} - 14]}{22} \\ M_e &= 25 \end{aligned}$$

Interpretation
This implies that 50% of the students got less than 25 marks and the other 50% got more than 25 marks.

The advantage of the median is that it is unaffected by outliers and is a useful measure of central tendency when the distribution of a random variable is severely skewed. A disadvantage of the median, however, is that it is inappropriate for categorical data. It is best suited as a central location measure for interval-scaled data such as rating scales.

3.6. Quartiles

Quartiles are those observations that divide an ordered data set into quarters (four equal parts). **Lower Quartile,** Q_1 is the first quartile or 25^{th} percentile. It is that observation which separates the lower 25 percent of the observations from the top 75 percent of ordered observations. **Middle Quartile,** Q_2 is the second quartile or 50^{th} percentile or the median. It divides an ordered data set into two equal halves. **Upper**

Downloaded by Khann (moin24126@gmail.com)

Quartile, Q_3 is the third quartile or 75^{th} percentile. It is that observation which divides observations into the lower 75 percent from the top 25 percent.

To compute quartiles, a similar procedure is used as for calculating median. The only difference lies in (i) the identification of the quartile position, and (ii) the choice of the appropriate quartile interval. Each quartile position is determined as follows:

For Q_1 use $\frac{n}{4}$, for Q_2 use $\frac{n}{2}$, and for Q_3 use $\frac{3n}{4}$ to calculate the position of the respective quartile. The appropriate quartile interval is that interval into which the quartile position falls. Like the median calculations, this is identified using the *less than ogive*. A forumular for Q_1 is:

$$Q_1 \quad = \quad O_{q1} + \frac{c(\frac{n}{4} - F(<))}{f_{q1}}$$

where O_{q1} is the lower limit of the class interval with lower quartile value, $F(<)$ is the cumulative frequency of the class interval before the lower quartile interval and f_{q1} is the frequency of the lower quartile interval.

3.6.1. Quartiles for ungrouped data

Ungrouped data situation is easy to calculate the quartiles. Simply identify the quartile positions and identify the value of the variable that lies at that position.

Exercise
Using income data below, find Q_1 and Q_2.

Income ($)	3800	4100	4400	4900	5200	5500	6000
Number of workers	12	13	25	17	15	12	6

Solution
n = 100, hence Q_1 position is at $\frac{n}{4} = \frac{100}{4} = 25^{th}$ position. Arranging number of workers cummulatively i.e. coming up with a cummulative distribution table 25^{th} value lies at income $4100. Hence Q_1 is $4100. Show that Q_3 is $5200.

3.6.2. Quartiles for grouped data

In calculating quartiles for grouped data, use of the formular is required since the position of the quartile will be a interval of numbers. The formular allows us to find

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

the exact value. Find the first, second and third quartile values from the distribution below.

Mark	0-9	10-19	20-29	30-39	40-49
Frequency	2	12	22	8	6

The lower quartile, Q_1
Q_1 position $\frac{n}{4} = \frac{50}{4} = 12.5^{th}$ position. Q_1 interval = [10 - 19] because the 12.5^{th} observation falls within these limits. The formula for Q_1 is:

$$Q_1 \quad = \quad O_{q1} + \frac{c(\frac{n}{4} - F(<))}{f_{q1}}$$

Where Q_1 is the lower quartile, O_{q1} is the lower limit of Q_1 Interval (class), n is the sample size (total number of observations), $F(<)$ is the cumulative frequency of the interval before the Q_1 interval, f_{q1} is the frequency of the Q_1 interval and c is the width of the Q_1 interval.

Thus:

$$Q_1 = O_{q1} + \frac{c[\frac{n}{4} - F(<)]}{f_{q1}} = 10 + \frac{10[\frac{50}{4} - 2]}{12} = 18.75$$

Interpretation:
25 % of the students got below 18.75 marks.

3.6.3. The second quartile, Q_2 (Median)

Q_2 position, use $\frac{n}{2} = \frac{50}{2} = 25^{th}$ position. Q_2 interval = [20 - 29] because the 25^{th} observation falls within these limits. The formula for Q_2 is:

$$Q_2 = O_{q2} + \frac{c[\frac{n}{2} - F(<)]}{f_{q2}}$$

Which gives Q_2 = 25 marks.

3.6.4. The upper quartile, Q_3

Q_3 position $\frac{3n}{4} = \frac{3\times 50}{4} = 37.5^{th}$ position. Q_3 interval = [30 - 39] because the 37.5^{th} observation falls within these limits. The formula for Q_3 is:

$$Q_3 = O_{q3} + \frac{c[\frac{3n}{4} - F(<)]}{f_{q3}}$$

where Q_3 is the upper quartile, O_{q3} is the lower limit of Q_3 class interval, n is the sample size (i.e. total number of observations), $F(<)$ is the cumulative frequency of the interval before the Q_3 interval, f_{q3} is the frequency of the Q_3 interval and c is the width of the Q_3 interval.

Downloaded by Khann (moin24126@gmail.com)

Thus:

$$Q_3 = O_{q3} + \frac{c[\frac{3n}{4} - F(<)]}{f_{q3}} = 30 + \frac{10[\frac{3\times 50}{4} - 36]}{8} = 31.875$$

Interpretation:
75% of the students got below 31.875 marks. Alternatively, 25% of the students got above 31.875 marks.

3.6.5. Percentiles

In general, any percentile value can be found by adjusting the median formula to: (i)Find the required percentiles position and from this and (ii) Establish the percentile interval.

Example
90^{th} percentile position = $0.9 \times n$, 35^{th} percentile position = $0.35 \times n$, 25^{th} percentile position(Q_1) = $0.25 \times n$

Uses of percentiles:
Percentiles are used to identify various non-central values. For example, if it is desired to work with a truncated dataset which excludes extreme values at either end of the ordered dataset.

3.7. Skewness

Skewness is departure from symmetry. Departure from symmetry is observed by comparing the mean, median and mode.

1. If $mean = median = mode$ the frequency distribution is Symmetrical.

2. If $mean < median < mode$ the frequency distribution is negatively skewed i.e. skewed to the left.

3. If $mean > median > mode$ the frequency distribution is positively skewed i.e. skewed to the right.

Remark:

1. If a distribution is distorted by extreme values (i.e. skewed) then the median or the mode is more representative than the mean.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

2. If the frequency distribution is skewed, the median may be the best measure of central location as it is not pulled by extreme values, nor is it as highly influenced by the frequency of occurrence.

3.8. Kurtosis

Kurtosis is the measure of the degree of peakedness of a distribution. Frequency distributions can be described as: leptokurtic, mesokurtic and platykurtic.

- Leptokurtic- highly peaked distribution i.e. a heavy concentration of observations of around the central location.
- Mesokurtic moderately peaked distribution.
- Platykurtic flat distribution i.e. the observations are widely spread about the central location.

3.9. Exercises

1. The number of days in a year that employees in a certain company were away from work due to illness is given in the following table:

Sick days	Number of employees
5-6	67
7-8	91
9-10	67
11-12	5

   Find the modal class and the modal sick days and interpret.

2. A company employs 12 persons in managerial positions. Their seniority (in years of service) and sex are listed below:

Sex	F	M	F	M	F	M	M	F	F	F	F	M
Seniority (yrs)	8	15	6	2	9	21	9	3	4	7	2	10

   (a) Find the seniority mean, median and mode for the above data.

   (b) Which of the mean, median and mode is the least useful measure of location for the seniority data? Give a reason for your answer.

   (c) Find the mode for the sex data. Does this indicate anything about the employment practice of the company when compared to the medians for the seniority data for males and females?

Downloaded by Khann (moin24126@gmail.com)

Chapter 4

Measures of Dispersion

4.1. Introduction

Spread or Dispersion refers to the extent by which the observations of a random variable are scattered about the central value. Measures of dispersion provide useful information with which the reliability of the central value may be judged. Widely dispersed observations indicate low reliability and less representativeness of the central value. Conversely, a high concentration of observation about the central value increases confidence in the reliability and representativeness of the central value.

4.2. Range

The range if the difference between the highest and the lowest observed values in a dataset.

For ungrouped dataset,

$$Range = X_{max} - X_{min}$$

For grouped dataset,

$$Range = Upper\ limit\ of\ last\ interval - Lower\ limit\ of\ first\ interval$$

The range is a crude estimate of spread. It is calculated, but is distorted by extreme values (outliers). An *outlier* would either be x_{max} or x_{min}. It is therefore a volatile and unstable measure of dispersion. It also provides no information on the clustering of observations within the dataset about a central value as it uses only two observations in its computation.

Example 6:
Given the following data in a frequency distribution table, find the range.
Solution:

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Income /$	3800	4100	4400	4900	5200	5500	6000
Number of Workers	12	13	25	17	15	12	6

$$Range = X_{max} - X_{min} = 6000 - 3800 = 2200$$

.

For grouped distribution with class intervals, x_{min} is the lower limit of the lower class interval and x_{max} is the upper limit of the highest class interval.

Interquartile range, IQR

Because the range can be distorted by extreme values, a modified range which excludes these outliers is often calculated. This modified range is the difference between the upper and lower quartiles i.e.

$$Interquartile\ Range = Q_3 - Q_1$$

This modified range removes some of the instability inherent in the range if outliers are present, but it excludes 50 percent of all observations from further analysis. This measure of dispersion, like the range, also provides no information on the clustering of observations within the dataset as it uses only two observations.

Quartile deviation

A measure of variation based on this modified range is called quartile deviation (QD) or the semi-interquartile range. It is found by dividing the interquartile range in half i.e.

$$Quartile\ deviation = \frac{Q_3 - Q_1}{2}$$

Remember when calculating this measure you order your dataset first to calculated Q_3 and Q_1. The quartile deviation is an appropriate measure of spread for the median. It identifies the range below and above the median within which 50 percent of observations are likely to fall. It is a useful measure of spread if the sample of observations contains excessive outliers as it ignores the top 25 percent and bottom 25 percent of the ranked observations.

4.3. Variance

The most useful and reliable measures of dispersion are those that take every observation into account and are based on an average deviation from a central value. Variance is such a measure of dispersion. Population variance is denoted by σ^2 whereas sample variance is denoted by s^2.

Downloaded by Khann (moin24126@gmail.com)

Variance for ungrouped data

Sample variance for ungrouped data is given by:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Population variance is given by:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The main difference being on the denominator of the two. Population variance divides by N whereas sample variance divide by $n-1$.

Consider the ages, in years, of 7 second hand cars: 13, 7, 10, 15, 12, 18, 9. Find the variance of the ages of cars.

Solution

Step 1: Find the sample mean, $\bar{x} = \frac{84}{7} = 12$ years.

Step 2: Find the squared deviation of each observation from the sample mean. See table below.

Car age, x_i	Mean, $\bar{x}$	Deviation $(x_i - \bar{x})$	Deviations squared $(x_i - \bar{x})^2$
13	12	+1	1
7	12	-5	25
10	12	-2	4
15	12	+3	9
12	12	0	0
18	12	+6	36
9	12	-3	9
		$\sum(x_i - \bar{x})$ = 0	$\sum(x_i - \bar{x})^2$ = 84

Step 3: Find the average squared deviation that is the variance using the formular:

$$S^2 = \frac{84}{7-1} = 14\ years^2$$

Note:

Divison by *n* would appear logical, but the variance statistic would then be a biased measure of dispersion. It can be shown to be unbiased if division is by $(n-1)$. For large samples i.e. n is greater than 30, however this distinction becomes less important.

Variance can be also calculated using a formular below. It gives similar results as the above formular:

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

$$S^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

$\sum x^2 = 1092$, $\sum x = 84$, n=7 and $\bar{x} = 12$, substituting the values in the above formular:

$$S^2 = \frac{[1092 - 7(12^2)]}{7-1} = \frac{84}{6} = 14\ years^2$$

Variance for grouped data

Grouped data is data presented in a frequency distribution table. Sample variance for such grouped data is calculated using the formular:

$$S^2 = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n-1}$$

or,

$$S^2 = \frac{\sum f x_i^2 - n\bar{x}^2}{n-1}$$

Population variance is given by:

$$\sigma^2 = \frac{\sum f x_i^2 - N\mu^2}{N}$$

Example 7:

Consider data for student marks obtained from Test 1. Calculate the variance of the student marks.

Marks	0-10	10-20	20-30	30-40	40-50
Frequency	2	12	22	8	6

Solution

The midpoint in an inteval is calculated as:

$$Midpoint = \frac{Lower\ limit + Upper\ limit}{2}$$

Marks, x	Frequency, f_i	Midpoint, x_i	$f_i x_i$	x^2	fx^2
0-10	2	$\frac{0+10}{2} = 5$	10	25	50
10-20	12	15	180	225	2700
20-30	22	25	550	625	13750
30-40	8	35	280	1225	9800
40-50	6	45	270	2025	12150
Total	50		1290		38450

Mean, $\bar{x} = \frac{\sum fx}{n} = \frac{1290}{50} = 25.8$ and $\sum fx^2 = 38450$. Using the above formular, the variance,

Downloaded by Khann (moin24126@gmail.com)

$$S^2 = \frac{\sum fx_i^2 - n\bar{x}^2}{n-1}$$

$$S^2 = \frac{38450 - 50(25.8)^2}{50-1} = \frac{5168}{49} = 105.47\ marks^2$$

The variance is a measure of average squared deviation about the arithmetic mean. It is expressed in squared units. Consequently, the meaning in a practical sense is obscure. To provide meaning, the measure should be expressed in the original units of the random variable.

4.4. Standard deviation

A standard deviation is a measure which expresses the average deviation about the mean in the original units of the random variable. The standard deviation is the square root of the variance. Mathematically:

A sample standard deviation is:

$$S_x = \sqrt{Sample\ variance} = \sqrt{s^2}$$

$$S_x = \sqrt{\frac{\sum fx_i^2 - n\bar{x}^2}{n-1}}$$

A population standard deviation is

$$\sigma^2 = \sqrt{\frac{\sum fx_i^2 - N\mu^2}{N}}$$

The standard deviation is a relatively stable measure of dispersion across different samples of the same random variable. It is therefore a rather powerful statistic. It describes how the observations are spread about the mean.

4.5. Coefficient of variation

From a sample, the coefficient of variation is defined as follows

$$CV = \frac{S}{\bar{x}} \times 100\%$$

whereas a population coefficient of variation is

$$CV = \frac{\sigma}{\mu} \times 100\%$$

.

This ratio describes how large the measure of dispersion is relative to the mean of

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

the observation. A coefficient of variation value close to zero indicates low variability and a tight clustering of observations about the mean. Conversely, a large coefficient of variation value indicates that observations are more spread out about their mean value.

From our example above,

$$CV = \frac{S}{\bar{x}} \times 100\% = \frac{10.27}{25.8} \times 100\% = 39.8\%.$$

4.6. Exercises

1. Find the mean and the standard deviation for the following data which records the duration of 20 telephone hotline calls on the 0772 line for advice on car repairs.

Duration	Number of calls
0 - ≤ 1	7
1 - ≤ 2	0
2 - ≤ 3	3
3 - ≤ 4	1
4 - ≤ 5	9

   At a cost of $2.60 per minute, what was the average cost of a call, and what was the total cost paid by the 20 telephone callers. Calculated the coefficient of variation and interpret it.

2. Employee bonuses earned by workers at a furniture factory in a recent month (US$) were:

47	31	42	33	58	51	25	28
62	29	65	46	51	30	43	72
73	37	29	39	53	61	52	35

   From the table above, find the

   (a) Mean and standard deviation of bonuses.

   (b) Interquartile range and quartile deviation.

   (c) Coefficient of variation and comment.

3. Give three reasons why the standard deviation is regarded as a better measure of dispersion than the range.

4. Discuss briefly which measure of dispersion would you use if the:

   (a) mean is used as the measure of central location and why?

Downloaded by Khann (moin24126@gmail.com)

(b) median is used as a measure of central location and why?

5. Discuss the limitations of the range as a measure of dispersion.

6. Define the following terms as they are used in statistics.

   (a) Outliers

   (b) Skewness

   (c) Kurtosis

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 5

Basic Probability

5.1. Introduction

This unit will introduce to you simple concepts and terminologies in probability. These include events, types of probabilities and rules of probabilities. Probability theory is fundamental to the area of statistical inference. Inferential statistics deals with generalising the behaviour of random variables from sample findings to the broader population. Probability theory is used to quantify the uncertainties involved in making these generalisations.

5.2. Definition

An Event is a collection of possible outcomes from an experiment or a trial. For example Heads or Tails are events which can be obtained from tossing a fair coin.

An Experiment is a process which generates events or outcomes. For instance, tossing a fair die three times constitutes an experiment.

Probability is the chance or likelihood of a particular outcome out of a number of possible outcomes occurring for a given event. Thus probability is a number between 0 and 1, which quantifies likelihood of an event occurring or not occurring. Therefore probability range is $0 \leq P(A) \leq 1$, where A is an event of a specific type.

Most decisions are made in the face of uncertainty. Probability is therefore, concerned with uncertainty.

5.3. Approches to probability theory

There are two broad approaches to probability, namely **Subjective** and **Objective**.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Subjective probability
It is probability which is based on a personal judgement that a given event will occur. There is no theoretical or empirical basis for producing subjective probabilities. In other words this probability of an event based on an educated guess, expert opinion or just plain intuition. Subjective probabilities cannot be statistically verified and there are not extensively used, hence will not be considered further.

Examples

1. When commuters board a commuter omnibus, they assume that they will arrive safely at their destinations, so P(arriving safely) = 1.

2. If you invest, you assume that you will get a good return, so P (good return) = 0.9.

Objective probabilities
These are probabilities that can be verified, through repeated experimentation or empirical observations. Mathematically it is defined as a ration of two numbers:

$$P(A) = \frac{r}{n}$$

Where A is event of a specific type, r is number of outcomes of event A, n is total number of possible outcomes (also called the sample space) and P(A) is probability of event A occurring.

Objective probabilities are derived either:

- **a priori** - that is when possible outcomes are known in advance such as tossing a coin,selecting cards from a deck of cards. Classical probability.

  $$P(A) = \frac{Number\ of\ outcomes\ favouring\ event\ A}{Total\ number\ of\ possible\ outcomes}$$

  For example the probability of a Head if a fair coin is tossed once is, $P(Head) = \frac{1}{2} = 0.5$

- **Empirically** that is when the values of r and n are not known in advance and have to be observed through data collection or from a relative frequency table you can deduce probability of the different outcomes.

  $$P(A) = \frac{Number\ of\ times\ event\ A\ has\ occured}{Number\ of\ times\ event\ A\ could\ have\ occured}$$

Downloaded by Khann (moin24126@gmail.com)

For instance, If out of a random sample of 90 customers 50 said they prefer Bakers Inn bread, then relative frequency that a randomly selected customer will prefer Bakers Inn bread is $\frac{50}{90} = 0.56$

- **Theoretically** - that is through use of theoretical distribution functions (mathematical formula that can be used to compute probabilities for certain event types).

These probabilities are used extensively in statistical analysis.

5.4. Properties of probability

1. A probability value lies only between 0 and 1 that is $0 \leq P(A) \leq 1$.
2. If an event A cannot occur (i.e. an impossible event), then $P(A) = 0$
3. If an event A is certain to occur, then $P(A) = 1$
4. The sum of probabilities of all possible outcomes of a random experiment equals one, that is $\sum_{i=1}^{n} P(E_i) = 1$.
5. Complementary probabilities: If $P(A)$ is the probability of event A occurring, then the probability of event A not occurring is defined as: $P(A^c) = 1P(A)$. Note: $P(A^c)$ is also sometimes written as $P(\bar{A}) or P(A')$.

Example
Consider random process of drawing cards from a card deck. These probabilities are called *a priori probabilities*.

1. Let A = event of selecting a *red card*. Then $P(Red\ card) = \frac{26}{52} = \frac{1}{2}$ (26 possible red cards out of 52 cards).
2. Let B = event of selecting a spade. Then $P(Spade) = \frac{13}{52} = \frac{1}{4}$ (13 possible spades out of a total of 52).
3. Let C = event of selecting an ace. Then $P(Ace) = \frac{4}{52} = \frac{1}{13}$ (4 possible ace out of a total of 52 cards).
4. Let D = event of selecting 'not an ace'. Then $P(not\ an\ ace) = 1P(ace) = 1 - \frac{1}{13} = \frac{12}{13}$.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

5.5. Basic probability concepts

1. **Intersection of Two events**
   The intersection of two events A and B is the set of outcomes that belongs to both A and B simultaneously. It is written as $A \cap B$ i.e. A and B and the keyword is *and*.

2. **Union of Two events**
   The union of Events A and B is the set of outcomes that belongs to either A or B or both and the key word is *or*. It is written as $A \cup B$.

3. **Complement of an event**
   The complement of an event A is the collection of all possible outcomes that are not contained in event A. That is $P(A^c) = 1P(A)$. Note $P(A^c)$ is also sometimes written as $P(\bar{A})$ or $P(A')$. In other words $P(A) + P(A') = 1$.

5.6. Types of events

1. **Mutually Exclusive or disjoint events**
   These are events which cannot occur at the same time. The occurrence of one event automatically prevents the occurrence of the other event. For mutually exclusive events the intersection of events is empty i.e. there are no common events.

   Examples

   (a) Passing and failing the same examination are mutually exclusive. In other words its not possible to pass and fail at the same time one examination.

   (b) In tossing a fair die once, getting a 3 and a 5 are mutually exclusive. You get one outcome at time and not both.

2. **Non-Mutually Exclusive Events**
   These are events which can occur simultaneously. The occurrence of one event does not prevent the occurrence of the other events. The intersection set of the events has got elements.

   Examples

   (a) In tossing a fair die once, getting an odd number or a number greater than 2 are non mutually exclusive events i.e. it is possible for the number to be odd and at the same time being greater than 2.

   (b) An individual can have more than one bank account i.e. if you open a bank account it does not prevent you from opening another account with another bank.

Downloaded by Khann (moin24126@gmail.com)

3. **Collectively exhaustive events**
   Events are said to be collectively exhaustive when the union of all possible events is equal to the sample space. This means that, in a single trial of a random experiment, at least one of these events is certain to occur.

   Example
   Consider a random experiment of selecting companies from the Zimbabwe Stock Exchange (ZSE). Let event A = small company, event B = medium company and event C = large company. Then $(A \cup B \cup C)$ = sample space (small, medium, large companies) = all ZSE companies.

4. **Statistically Independent events**
   Two events are said to statistically independent if the occurrence of event A has no effect on the outcome of event B occurring and vice versa.

   Example
   Let A = event that an employee is over 30 years of age, B = event that the employee is female. If it can be assumed or empirically verified that a randomly selected employee is over 30 years of age from a large organisation is equally likely to be either male or female employee, then the two events A and B are statistically independent.

5. **Statistically Dependent events**
   Events are dependent if the occurrence of one of the event A affects the occurrence of the second event B. These will be discussed under conditional probability.

   Remark
   The terms Statistically independent events and mutually exclusive events should not be confused. They are two very different concepts. When two events are mutually exclusive, they are NOT Statistically independent. They are dependent in the sense that if one event happens, then the other event cannot happen. In probability terms, the probability of the intersection of two mutually exclusive events is zero, while the probability of two independent events is equal to the product of the probabilities of the separate events.

5.7. Laws of probability

There are generally two laws in probability theory, namely, Addition Laws and Multiplication Laws. **Addition laws** pertain to mutually and non-mutually exclusive events only. The key word is **OR**. What is the probability that event A **OR** B will occur:

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

- **For Mutually Exclusive events:**
  $P(A \textbf{or} B) = P(A) + P(B)$
  $P(A \cup B) = P(A) + P(B)$
  Note: Union ($\cup$) means OR

  Example
  What is the probability of getting a 5 or 6 if a fair die is tossed once?

  Solution The sample space has six possible outcomes $1, 2, 3, 4, 5, 6$. Therefore $P(5 or 6) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

- **For Non-mutually Exclusive events**
  $P(A \textbf{ or } B) = P(A) + P(B) - P(A \; and \; B)$
  $P(A \cup B) = P(A) + P(B) P(A \cap B)$

  The intersection sign, $\cap$ means the joint probability of events A and B. $P(A \; and \; B)$ is subtracted to avoid double counting.

  Example
  What is the probability of getting an even number or a number less than four if a fair die is tossed once?

  Answer
  Let event A = getting an even number and the elements are 2, 4, 6 and event B = getting a number less than four and the elements are 1, 2, 3. Then $P(A) = \frac{3}{6} \; and \; P(B) = \frac{3}{6}$. Thus $P(A \; and \; B) = \frac{1}{6}$. There is only one element which is common in A and B and the number is 2. Therefore
  $P(A \; or \; B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}$

Exercise 1
Sixty per cent of the population of a town read either magazine A or magazine B and 10% read both. If 50% read magazine A, what is the probability that one person, selected at random, read magazine B?

Multiplication Laws
Multiplication Laws pertain to dependent and independent events. The key word is **AND**

a) For Independent events: $P(A \; and \; B) = P(A) \times P(B)$

Downloaded by Khann (moin24126@gmail.com)

Example
What is the probability of getting a tail when two fair coins are tossed at the same time?

Answer
$P(T1\ and\ T2) = P(T1) \times P(T2)$
T1 = the probability of getting a tail from coin 1. T2 = the probability of getting a tail from coin 2. The two outcomes do not affect each other. Therefore
$P(T1\ and\ T2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

b) Dependent events will be discussed on the section on conditional probability.

5.8. Types of probabilities

Objective probabilities can be classified into 3 categories, namely:

- Marginal Probability
- Joint Probability
- Conditional probabilities

Marginal Probability
It is the probability of only a single event A occurring regardless of certain conditions prevailing. It is written as P (A). A frequency distribution describes the occurrence of only one characteristic of interest at a time and is used to estimate marginal probabilities.

Joint Probability
It is the chance that two or more events will occur simultaneously. It is the occurrence of events at the same time. If the joint probability on any two events is zero, then the events are mutually exclusive.

Conditional Probability
It is the probability that a given event occur under the conditions that another event has already occurred. The symbol $P(A|B)$ is the probability that the event A will occur given that event B has already occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{5.1}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Payment Method	Sex		Total
	Male	Female	
Credit Card	10	15	25
Cash	8	6	14
Total	18	21	39

provided $P(B) > 0$, and similarly

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{5.2}$$

provided $P(A) > 0$

Note: $P(A \cap B) = P(B|A).P(A) = P(A|B).P(B)$. P(A and B) is the joint probability of events A and B. $P(B)$ is the probability of event B, which is a marginal probability.

Example: Joint Marginal and Conditional Probabilities

Consider the table of fees payment methods by sex

What is the probability of getting a person who is

a) (i) Female and uses a credit card?

(ii) Male and uses cash?

b) (i) Credit card user?

(ii) Female?

c) (i) Female given that she uses cash?

(ii) Credit card user given that he is a male?

Answer

(a) Question a) is joint probability of the events. The sample space has 39 people altogether.
(i) P(female and credit card) = $\frac{15}{39}$ = 0.3846.

Note: The two events should not be confused with independent events. In this case find the value in the intersection set of female column and credit card row which is 15.

(ii) P(male and cash) = $\frac{8}{39}$ = 0.2051.
It is the chance of two events occurring at the same time.

(b) Question b) is marginal probability, the conditions prevailing is sex or payment method.

Downloaded by Khann (moin24126@gmail.com)

(i) P(credit card user) = $\frac{25}{39}$ = 0.6.

Note: The prevailing condition which has been ignored is sex.
(ii) P(female) = $\frac{21}{39}$ = 0.5385.

The condition which has been ignored is payment method. For joint probabilities, consider values inside the table as a ration of the grand total 39. For marginal probabilities consider row and column totals as ratios of grand total.

c) Question c) is conditional probability.

i) P(female — cash user) $= \frac{P(Female\ and\ Cash)}{P(Cash)} = \frac{\frac{6}{39}}{\frac{14}{39}} = \frac{6}{14} = 0.4286$

ii) P(credit card — male)$= \frac{P(Credit\ card\ and\ Male)}{P(Male)} = \frac{\frac{10}{39}}{\frac{18}{39}} = \frac{5}{9} = 0.5556$

Exercise 2
A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others are blue. He gets dressed in the dark, so he just grabs a shirt and puts it on. He plays golf two days in a row and does not do laundry. What is the likelihood both shirts selected are white?

5.9. Contigency Tables

A contingency table is a cross-tabulation that simultaneously summarise two variables of interest and their relationship. The level of measurement can be nominal. It is a table that is used to classify sample observations according to two or more identifiable characteristics.

Example
A survey of 150 students classified each as to gender and the number of movies attended last month. Each respondent is classified according to two criteria, that is, the number of movies attended and gender.

Movies Attended	Gender		Total
	Male	female	
0	20	40	60
1	40	30	70
2 or more	10	10	20
Total	70	80	150

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

5.10. Tree diagram

A tree diagram is a graph that is helpful in organising calculations that involves several stages. Each branch in the tree is one stage of the problem. The branches of the tree diagram are weighted by probabilities.

Exercise 3
Using the previous example of 150 students? Using a tree diagram what is the probability of selecting a male student given that he has seen one movie?

5.11. Counting rules

Probability computations involve counting the number of successful outcomes (r) and the total number of possible outcomes (n) and expressing them as a ratio. Often the values of rand n are not feasible to count because of the large number of possible outcomes involved. Counting rules assist in finding values of r and n. There are three basic counting rules:

- Multiplication rule
- Permutations
- Combinations.

5.11.1. Multiplication Rule

The multiplication rule is applied in two ways:

a) The total number of ways in which n objects can be arranged (ordered) is given by:
$n! = nfactorial = n(n-1)(n-2)(n-3)$.....3.2.1
Note that 0! = 1.

Example
The number of different ways in which 7 horses can complete a race is given by:
$7! = 7.6.5.4.3.2.1 = 5040$ different arrangements.

b) If a particular random process has

- n_1 possible outcomes on the first trial
- n_2 possible outcomes on the second trial e.t.c.
- n_j possible outcomes on the first trial.

Downloaded by Khann (moin24126@gmail.com)

Then the total number of outcomes for the j trials is: $n_1 \times n_2 \times n_3 \times n_4 \times ..... \times n_j$

Example

A restaurant menu has a choice of 4 starters, 10 main courses and 6 desserts. What is the total number of meals that can be ordered in this restaurant.

Solution

The total numbers of possible meals that can be ordered are: $4 \times 10 \times 6 = 240$ meals.

5.11.2. Permutations

A permutation is number of distinct ways in which a group of objects can be arranged. Each possible arrangement (ordering) is called a permutation. The number of ways of arranging r objects selected from n objects where ordering is important, is given by the formula:

$$P_r^n = \frac{n!}{(n-r)!} \tag{5.3}$$

Where $n! = n\ factorial = n(n-1)(n-2)(n-3)....3.2.1.$ and r = numbers of objects selected at a time and n = total number of objects.

Example

10 horses compete in a race.

(i) How many distinct arrangements are there of the first 3 horses past the post?

(ii) What is the probability of predicting the order of the first 3 horses past the post?

Answer

(i) Since the order of 3 horses is important, it is appropriate to use the permutation formula.
That is: $P_r^n = P_3^{10} = \frac{10!}{(10-3)!} = 720$
There are 720 distinct ways of selecting 3 horses out of 10 horses.

(ii) The probability of selecting the first 3 horses past the post is:

$P(first\ 3\ horses) = \frac{1}{Selecting\ 3\ out\ of\ 10\ horses} = \frac{1}{730}$ chance of winning.

5.11.3. Combinations

A combination is the number of different ways of arranging a subset of objects selected from a group of objects where the ordering is not important. Each possible arrangement is called a combination. The number of ways of arranging r objects selected from

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

n objects, not considering order, is given by the formula:

$$C_r^n = \frac{n!}{(n-r)!r!} \tag{5.4}$$

Where $n! = n\ factorial = n(n-1)(n-2)(n-3)...3.2.1$, $r! = r(r-1)(r-2)(r-3)...3.2.1$, r = number of objects selected and n = total number of objects.

Example
10 horses complete in a race.

(i) How many arrangements are there of the first 3 horses past the post, not considering the order in which the first three pass the post?

(ii) What is the probability of predicting the first 3 horses past the post, in any order?

Answer

(i) The order of the first 3 horses is not important, hence apply the combination formula.
$C_r^n = \frac{n!}{(n-r)!r!} = \frac{10!}{(10-3)!7!} = 120$
There are 120 different ways of selecting the first 3 horses out of 10 horses, without regard to order.

(ii) The probability of selecting the first 3 horses past the post, disregarding order is given by
$P(first\ 3\ horses) = \frac{1}{Selecting\ 3\ horses} = \frac{1}{120}$ chance of winning.

5.12. Exercise

1 . Find the values of:

(a) P_4^7

(b) C_2^8

2 . There are 5 levels of shelving in a supermarket. If 3 brands of soup must each be placed on a separate shelf, how many different ways can a packer arrange the soup brands?

3 . In an examination a student is asked to answer three questions from an examination paper containing eight questions. How many deferent selections are possible?

4 . A company has 12 products in its product range. It wishes to advertise in the local newspaper, but due to space constraints, it is only allowed to display 7 of

Downloaded by Khann (moin24126@gmail.com)

its products at a time. How many different ways can this company compose a display in the local newspaper?

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 6

Probability Distributions

6.1. Introduction

This unit will study probability distributions. A probability distribution gives the entire range of values that can occur based on an experiment. A probability distribution is similar to a relative frequency distribution. However in steady of describing the past, it describes a likely future event. For instance a drug manufacturer may claim a treatment will cause weight loss for 80% of the population. A consumer protection agency may test the treatment on a sample of six people. If the manufacturers claim is true, it is almost impossible to have an outcome where no one in the sample loses weight and its most likely that 5 out of the 6 do lose weight.

6.2. Definition

A probability distribution or probability mass function is a listing of all the possible outcomes of an experiment and the probability of each of these outcomes.

6.3. Random variables

A random variable is a function whose value is a real number determined by each element in the sample space. In other words its a quantity resulting from an experiment that, by chance, can assume different values. There are two types of random variables, discrete random variables and continuous random variables.

Discrete Random Variable is a variable that can assume a countable number of values. Examples include:

- Number of defective light bulbs obtained when three light bulbs are selected at random from a consignment could be 0, 1, 2, or 3.
- The number of employees absent from the day shift on Monday.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

- The daily number of accidents that occur in the city of Harare.

Continuous Random Variable is a variable that can assume values corresponding to any of the points contained in one or more intervals. Examples include:

- The waiting time for customers to receive their order at a manufacturing company.
- Tire pressure measure in kilo Pascal (KPa) of an automobile.
- The height of each student is this class.

The choice of a particular probability distribution function depends primarily on the nature of the random variable (i.e. discrete or continuous) under study. Thus we have **Discrete** Probability distributions and **Continuous** Probability distributions.

6.4. Discrete probability distribution

The probability distribution or probability mass function of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible value the random variable can assume.

Example
Find the probability mass function of the numbers when a pair of dice is thrown.

Answer
Let X be a random variable whose values of x are the possible totals of the outcomes of the two dies. Then x can be an integer from 2 to 12. Two dice can fall in 6×6 ways each with a probability $\frac{1}{36}$. For example, $P(X = 3) = \frac{2}{36}$ since a total of 3 can occur in two ways, that is (1,2) or (2, 1). The probability distribution (mass) function is:

x	2	3	4	5	6	7	8	9	10	11	12
P(X= x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Exercise 1

1 Three coins are tossed. Let X be the number of heads obtained. Find the function of X.

2 Suppose you are interested in the number of Tails showing face up of coin. What is the probability distribution for the number of Tails?

Downloaded by Khann (moin24126@gmail.com)

6.5. Properties of discrete probability mass function

Let X be a discrete random variable that can assume values $x_1, x_2, x_3, \ldots, x_n$. Then $P(X = x)$ is the probability mass function if:

i. $0 \leq P(X = x) \leq 1$ for all values of x.

ii. $P(X = x_1) + P(X = x_2) + \ldots + P(X = x_n) = \sum_{i=1}^{n} = 1$

This means that probability of any value of x is never negative and also the sum of the probabilities of the discrete random variable should be equal to 1.

iii. The mean, or the expected value of a discrete random variable X is $\mu =$ and given by:

$$E(X) = \sum_{all\ x} x_i P(X = x_i)$$

iv. Variance of a discrete random variable X is

$$\begin{aligned} Var(X) &= \sigma^2 = E(X - \mu)^2 \\ \sigma^2 &= \sum_{all\ x} (x_i - \mu)^2 P(X = x_i) \end{aligned} \tag{6.1}$$

Note that: $\sigma^2 = E(X^2) - E(X)^2 = E(X^2) - \mu^2$

Example

Consider the following probability distribution for a discrete random variable. Verify

x	0	1	2	5	10
$P(X = x_i$	0.05	0.25	0.30	0.20	0.20

the probability properties and find the standard deviation of the distribution.

Solution

i. All $P(X = x_i)$ are between 0 and 1, for i = 0, 1, 2, 5, 10.

ii. Sum of probabilities should be equal to 1.

$$\begin{aligned} \sum P(X = x_i) &= P(X = x_1) + P(X = x_2) + \ldots + P(X = x_5) \\ &= 0.05 + 0.25 + 0.30 + 0.20 + 0.20 \\ &= 1 \end{aligned}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

iii. The mean is given by:

$$\begin{aligned}\mu &= \sum x_i P(X = x_i) \\ &= 0 \times 0.05 + 1 \times 0.25 + 2 \times 0.30 + 5 \times 0.20 + 10 \times 0.20 \\ &= 3.85\end{aligned}$$

iv. The variance is given by

$$\begin{aligned}Var(X) &= E(X-\mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) \\ &= (0 - 3.85)^2 \times 0.05 + (1 - 3.85)^2 \times 0.25 + \ldots + (10 - 3.85)^2 \times 0.20. \\ &= 11.6275\end{aligned}$$

v. Standard deviation, $\sigma = \sqrt{Var(X)} = \sqrt{11.6275} = 3.410$

6.6. Probability terminology and notation

a) At most 3, means not more than 3. Here 3 is an arbitrary number, it therefore means 3 is the maximum discrete value which can be assumed by a random variable. Let X be the random variable, taking $x = 0, 1, 2, 3, \ldots, n$, where n is the sample size. Notation: $P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$.

b) At least 3, means not less than 3. The minimum that can be assumed is 3 since 3 is not less than itself. Notation: $P(X \geq 3) = P(X = 3) + P(X = 4) + P(X = 5) + \ldots + P(X = n)$

c) Less than 3 this effectively means values below 3, and 3 is actually excluded. Notation: $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$

d) More than 3, means values above 3, in discrete terms it is from 4 upwards. Notation: $P(X > 3) = P(X = 4) + P(X = 5) + P(X = 6) + \ldots + P(X = n)$ or using the complimentary rule it is given as $1 - P(X \leq 3)$

e) Exactly 3, it means equals. Notation: $P(X = 3)$.

f) Between 3 and 6 means the discrete values between 3 and 6, which are 4 and 5. However it should be noted that the limits can be exclusive or inclusive. Notation for exclusive: $P(3 < X < 6) = P(X = 4) + P(X = 5)$. Notation for inclusive: $P(3 \leq X \leq 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$

Exercise 2

Consider the following probability distribution that characterises a marketing analysts belief concerning the probabilities associated with the number, x of sales that a company might expect per month for a new super computer:

Downloaded by Khann (moin24126@gmail.com)

x	0	1	2	3	4	5	6	7	8
P(X=x)	0.02	0.08	0.15	0.19	0.24	0.17	0.10	0.04	0.01

i. What is the probability that the company will sell as in (a), (b), (c), (d), (e) and (f) above.

ii. Find the mean, variance and standard deviation of X.

6.7. Discrete probability distributions

For the purpose of this course we will focus on three special and commonly used discrete probability distributions that are Bernoulli, Binomial and Poisson distributions.

6.7.1. Bernoulli distribution

In probability theory and statistics, the Bernoulli distribution, named after Swiss scientist Jacob Bernoulli, is the probability distribution of a random variable which takes value 1 with success probability and value 0 with failure probability. A random variable X which has two possible outcomes, say 0 and 1, is called a Bernoulli random variable. The probability distribution of X is:

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

i.e. $P(X = 0) = 1 - P(X = 1) = 1 - p$

This distribution best describes all situations where a ”trial” is made resulting in either ”success” or ”failure,” such as when tossing a coin, or when modeling the success or failure of a surgical procedure. The Bernoulli distribution is defined as:

$$f(X = x) = p^x(1 - p)^{1-x}, \quad \text{for } x = 0, 1 \tag{6.2}$$

Where, p is the probability that a particular event (e.g. success) will occur.

Example: tossing a fair coin, you get a head or a tail each with probability of 0.5. Thus, if a head is labelled 1 and a tail 0, and a tail the random variable X representing the outcome takes values 0 or 1. If the probability that $X = 1$ is p, then we have that:

$$P(X = 1) = \frac{1}{2}.$$

$$P(X = 0) = 1 - \frac{1}{2} = \frac{1}{2},$$

since events $X = 1$ and $X = 0$ are mutually exclusive.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

6.7.2. Binomial distribution

Data often arise in the form of counts or proportions which are realizations of a discrete random variable. A common situation is to record how many times an event occurs in n repetitions of an experiment, i.e. for each repetition the event either occurs (a "success") or it does not (a "failure") occur. More specifically, consider the following experimental process:

1. There are n trials.
2. Each trial results in a success or a failure.
3. The probability of a success, p is constant from trial to trial.
4. The trials are independent.

An experiment satisfying these four conditions is called a binomial experiment. The outcome of this type of experiment is the number of successes, i.e., a count. The discrete variable X representing the number of successes is called a binomial random variable. The possible counts, $X = 0, 1, 2, \ldots n$, and their associated probabilities define the binomial distribution, denoted by $Bin(n, p)$.

Suppose we repeat a Bernoulli p experiment n times and count the number X of successes, the distribution of X is called the Binomial, $Bin(n, p)$ random variable. The quantities n and p are called parameters and they specify the distribution.

If $X = X_1 + X_2 + \ldots + X_n$, where X_i are independent and identically distributed Bernoulli random variables, then X is called a Binomial random variable. Thus the probability mass function is:

$$P(X = x) = C_x^n p^x (1 - p)^{n-x} \tag{6.3}$$

Where $k = 0, 1, 2, \ldots, n$, $0 < p < 1$ and $C_x^n = \frac{n!}{(n-x)!x!}$

Mean and variance of the Binomial Distribution

1. Mean, $\mu_x = E(X) = np$
2. Variance, $\sigma_x^2 = Var(X) = np(1 - p) = npq$

Example:
The notation $Bin(n, p)$ means a Binomial distribution with parameters n and p. Find:

Downloaded by Khann (moin24126@gmail.com)

i) Probability of getting 4 heads in 6 tosses of a fair coin

ii) Mean and

iii) Variance.

Solution:

i) Let X is the number of heads (success) when a fair coin is tossed. Thus: $P(X = 4) = C_4^6 \frac{1}{2}^4 (1 - \frac{1}{2})^{64} = \frac{15}{64}$

ii) Mean$= np = 6 \times \frac{1}{2} = 3$

iii) Variance$= npq = 6 \times \frac{1}{2} \times (1 - \frac{1}{2}) = 1.5$

Exercises:

1. A manufacturer of nails claims that only 3% of its nails are defective. A random sample of 24 nails is selected, what is the probability that 5 of the nails are defective?

2. A certain rare blood type can be found in only 0.05% of people. If the population of a randomly selected group is 3000, what is the probability that at least two persons in the group have this rare blood type?

6.7.3. Poisson distribution

The Poisson distribution, named after French mathematician Simon Denis Poisson, is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

Examples binomial experiments

i) The number of cars arriving at a parking garage in one-hour time interval.

ii) The number of defective screws per consignment.

iii) Number of typing errors per page.

iv) Number of particles of a given chemical in a litre of water.

The Poisson question: What is the probability of r occurrences of a given outcome being observed in a predetermined time, space or volume interval?

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

A Poison random variable is a discrete random variable that can take integer values from 0 up to infinite (∞). The parameter for this distribution is λ, i.e. $Po(\lambda)$. The Poisson probability mass function is given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{6.4}$$

Where $x = 0, 1, 2, \ldots \infty$ and $0 < \lambda < \infty$.

Example:
The number of students arriving at a takeaway every 15 minutes is a Poisson random variable with parameter $\lambda = 0.2$. Find the probability that zero, at most one and more than two students arrive at the takeaway.

Solution:
Using the formular that:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

i) Probability that no students arrive:

$$\begin{aligned} P(X = 0) &= \frac{0.2^0 e{-}0.2}{0!} \\ &= 0.8187 \end{aligned}$$

ii) Probability that at most one student arrive

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1) \\ &= \frac{0.2^0 e{-}0.2}{0!} + \frac{0.2^1 e{-}0.2}{1!} \\ &= 0.9824 \end{aligned}$$

iii) Probability that at least two students arrive)

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + \ldots \\ &= 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - (0.8187 + 0.1637) \\ &= 0.0176 \end{aligned}$$

Properties of Poisson Random variable

Downloaded by Khann (moin24126@gmail.com)

1. Mean = $E(X) = \lambda$
2. Variance = $Var(X) = \lambda$

Exercises

1. A textile producer has established that a spinning machine stops randomly due to thread breakages at an average rate of 5 stoppages per hour. What is the probability that in a given hour:
   i) 3 stoppages will occur on this spinning machine?
   ii) at most 2 stoppages will occur on this spinning machine?
   iii) more than 4 stoppages will occur on this spinning machine?
   iv) between 2 and 6 stoppages will occur on this spinning machine?
   v) What is the probability that no more than 1 stoppage will occur in a given two-hour interval?
2. The arrival of patients at a rural clinic is 2 per hour. In any given hour what is the probability that:
   a) no patient will arrive?
   b) exactly six patients will arrive?
   c) not less than 2 patients will arrive.
   d) Determine the variance.

Remark: As a general rule, always check that the time, space or volume interval over which occurrences of the random variable are observed is the same as the time, space or volume interval corresponding to the average rate of occurrences, λ. When they differ, adjust the rate of occurrences to coincide with the observed interval.

6.8. Continuous probability distributions

We will discuss three continuous random variable distributions that are the Normal, Uniform and Exponential distributions only.

6.8.1. The Normal distribution

One of the most useful and frequently encountered continuous random variable distributions is called the Normal distribution. Its graph is called the normal curve, which is bell shaped. The curve describes the distribution of so many sets of data that occur in nature, industry and research.

Characteristics of the Normal Distribution

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

i. It is bell-shaped.

ii. It is symmetrical about a central value μ.

iii. The tails of the distribution never touch the axis (i.e. asymptotic).

iv. A normally distributed random variable if fully described by two parameters namely μ (the population mean) and σ^2 (the population variance).

v. The area under the curve is equal to 1.

A random variable, X is said to be normally distributed with a probability density function given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}, \qquad -\infty < x < \infty \tag{6.5}$$

where μ = mean of the random variable X and σ^2 = variance of the random variable X. The random variable X is represented as: $X \approx N(\mu, \sigma^2)$. μ and σ^2 are said to be the parameters of X.

It is difficult to use the probability density function of the normal distribution to calculate the probabilities for X. Hence the process of standardisation is used so that the probability values are taken directly from the standard nornal distribution table. This table indicates the probabilities corresponding to different values of Z starting at -3. The process of standardisation involves calculating the value of Z by use of a formular:

$$Z = \frac{X - \mu}{\sigma} \tag{6.6}$$

6.8.2. The standard normal distribution

The standard normal distribution is a special kind of a normal distribution with mean zero and variance 1. **Z** is called a standard normal distribution random variable and written as: $Z \approx N(0, 1)$. The cumulative distribution function of **Z** is denoted by $\Phi(Z)$, and is denoted by $P(Z < z) = \Phi(Z)$. The values of this are found in the Standard normal distribution tables.

Use Standard normal distributiion tables to find the probabilities below.

a. $P(Z \geq -2)$

b. $P(Z > 0.79)$

c. $P(-1.11 < Z < -0.7)$

d. $P(-1.3 < Z < 2.1)$

Downloaded by Khann (moin24126@gmail.com)

e. $P(Z \leq -3)$

f. $P(0.04 < Z < 0.46)$

Solution

a.

$$\begin{aligned} P(Z \geq -2) &= 1 - P(Z \leq -2) \\ &= 1 - \Phi(-2) \\ &= 1 - 0.0228 \\ &= 0.9772 \end{aligned}$$

b.

$$\begin{aligned} P(Z > 0.79) &= 1 - P(Z < 0.79) \\ &= 1 - \Phi(0.79) \\ &= 1 - 0.7852 \\ &= 0.2148 \end{aligned}$$

c.

$$\begin{aligned} P(-1.11 < Z < -0.7) &= \Phi(-0.79) - \Phi(-1.11) \\ &= 0.2420 - 0.1335 \\ &= 0.1085 \end{aligned}$$

d.

$$\begin{aligned} P(-1.3 < Z < 2.1) &= \Phi(2.1) - \Phi(-1.3) \\ &= 09821 - 0.0968 \\ &= 0.8853 \end{aligned}$$

e.

$$\begin{aligned} P(Z \leq -3) &= \Phi(-3) \\ &= 0.0013 \end{aligned}$$

f.

$$\begin{aligned} P(0.04 < Z < 1.46) &= \Phi(1.46) - \Phi(0.04) \\ &= 0.9278 - 0.5160 \end{aligned}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

$$= \quad 0.4118$$

Finding probabilities using Z-Distribution

The normal random variable, **X** with a mean μ and variance σ ($N(\mu, \sigma)$) has a standard normal distribution

$$z = \frac{x - \mu}{\sigma}$$

6.8.3. The Uniform distribution

This distribution is also known as the rectangular distribution. A continuous uniform variable has a probability density over an interval. Its probability distribution density function is:

$$f(X = x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & elsewhere \end{cases} \tag{6.7}$$

Properties of the Uniform distribution

i. Mean is given by

$$E(X) = \frac{1}{2}(b + a) \tag{6.8}$$

ii. Variance is given by

$$Var(X) = \frac{(b - a)^2}{12} \tag{6.9}$$

NB

The probability that X falls in some interval say (c,d) can be easily calculated by integrating the density function

$$f(x) = \frac{1}{b - a}$$

to obtain

$$P(X = (c, d)) = \frac{c - d}{b - a}$$

Example:

The marks of students from a certain examination are uniformly distributed in the interval 50 to 75. The density function for the marks is given by:

$$f(X = x) = \begin{cases} \frac{1}{75-50} & 75 < x < 50 \\ 0 & elsewhere \end{cases}$$

Find the mean and variance of this distribution.

Solution:

Downloaded by Khann (moin24126@gmail.com)

1. The Mean is given by $E(X) = \frac{1}{2}(b+a) = \frac{1}{2}(75+50) = 62.5$

2. The Variance is given by $\frac{(b-a)^2}{12} = \frac{(75-50)^2}{12} = 52.083$

Interpretation:
The average mark for the examination was 62.5 with a variance of 52.083.

Exercise:
For the continuous uniform distribution defined of the interval [a, b], where $b > a$, show that
i) Mean $= \frac{1}{2}(b+a)$ and
ii) Variance $= \frac{(b-a)^2}{12}$

6.8.4. The Exponential distribution

An exponential distribution variable is a continuous random variable that can take on any positive value. **X** is said to be an exponential random variable with a parameter λ and a probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & otherwise \end{cases} \tag{6.10}$$

The exponential distribution function ogten arises in practice as the distribution of waiting time i.e. the amount of time until a specified event occurs. Examples include time until a customer arrives, or time until a machine fails.

Example
Suppose that the length of a phone call in minutes is exponentially distributed with parameter, $\lambda = 0.1$. If someone arrives immediately ahead of you at a public telephone booth, what is the probability that you will wait for at least 20 minutes.

Solution
Let **X** be the length of a phoe call made in front of you. Then

$$\begin{aligned} P(X > 20) &= \int_{20}^{\infty} 0.1e^{-0.1x}dx \\ &= -e^{-0.1x}|_{20}^{\infty} \\ &= e^{-2} = 0.2706 \end{aligned}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 7

Interval Estimation

7.1. Introduction

We are now in the knowledge that a population parameter can be estimated from sample data by calculating the corresponding point estimate. This chapter is motivated by the desire to understand the goodness of such a point estimate. However, due to sampling variability, it is almost never the case that the population parameter equals the sample statistic. Further, the point estimate does not provide any information about its closeness to the true population parameter. Thus, we cannot rely on point estimates for decision making and policy formulation in day to day living and or in any organisation, institution or country. We need bounds that represent a range of plausible values for a population parameter. Such ranges are called ***interval estimates***.

To obtain the interval estimates, the same data from which the point estimate was obtained is used. Interval estimates may be in the form of a *confidence interval* whose purpose is to bound population parameters such as the mean, the proportion, the variance, and the standard deviation; a *tolerance interval* which bounds a selected proportion of the population; and a *prediction interval* which places bounds on one or more future observations from a population.

In this course, we will focus on confidence intervals.

7.2. Confidence Intervals

It is noted that we cannot be certain that an interval contains the true but unknown population parameter since only a sample from the full population is used to compute both the point estimate and, the interval estimate! A confidence interval is constructed so that there is high confidence that it does contain the true but unknown population parameter. Generally, a $100(1-\alpha)\%$ confidence interval equals

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

point estimate $\pm$ *reliability coefficient* $\times$ *s.e.(parameter)*

where α is the *level of significance* between zero and one; $1-\alpha$ is a value called the *"confidence coefficient"*; $100(1-\alpha)\%$ is the *confidence level*; *parameter estimate* is a value for the point estimate such as for the sample mean, $\overline{x}$, or for the population proportion, $\widehat{p}$; *reliability coefficient* is a probability point obtained from an appropriate table as dictated by, for example, z_α or $t_{\frac{\alpha}{2},n-1}$; and *s.e.(parameter)* read *standard error of the parameter*, measures the closeness of the point estimate to the true population parameter i.e. it measures the precision of an estimate in getting the parameter.

7.3. Confidence Interval for the Population Mean

The overall assumption made is that the sample comes from a normal population.

Case 1: *Known Population Variance*
Suppose that, in addition to the overall assumption, the variance of the population, σ^2, is known. Then a random variable called the *sample mean*, $\overline{X}$, is defined such that

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}),$$

whose standardised result is

$$Z = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1).$$

The $100(1-\alpha)\%$ confidence interval estimate for the population mean may also take the form $\ell_1 \leq \mu \leq \ell_2$ where the end points ℓ_1 and ℓ_2 are called lower- and upper-confidence limits respectively and are computed from the sample data. Different samples will produce different values for the end points.

Also, observe that,

$$\ell_1 = \overline{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

and

$$\ell_2 = \overline{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

Thus, a $100(1-\alpha)\%$ confidence interval for the population mean is

$$\overline{x} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \leq \ \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}.$$

Example

Downloaded by Khann (moin24126@gmail.com)

Consider the following data.

64.3 64.6 64.8 64.2 64.5 64.3 64.6 64.8 64.2 64.3

Assume that it is normally distributed with unit population variance. For these data, construct a 95% confidence interval for the population mean.

Solution
Using the data, $n = 10$, $\overline{x} = 64.46$, the level of significance, $\alpha = 5\% = 0.05$, and from the given assumption, $\sigma^2 = 1$. Now, the resulting 95% confidence interval (CI) for the population mean is

$$\overline{x} - z_{0.025} \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{0.025} \times \frac{\sigma}{\sqrt{n}}$$

Substituting we have

$$64.46 - 1.96 \times \frac{1}{\sqrt{10}} \leq \mu \leq 64.46 + 1.96 \times \frac{1}{\sqrt{10}}.$$

Simplifying we then have the 95% CI for the population mean as

$$63.84 \leq \mu \leq 65.08.$$

Interpretation
From the above estimation of the confidence interval for the population mean, it is tempting to conclude that the population mean, μ, is within $63.84 \leq \mu \leq 65.08$ with a probability of 0.95. To be blunt, this statement is not true.

Well, the true value of the population mean, μ, is unknown (!), and the confidence interval is a random interval that is a function of the sample mean (!). In this scenario, to say that μ is within $63.84 \leq \mu \leq 65.08$ with a probability 0.95 is *totally* off the mark. Now, if you follow the argument, the statement $63.84 \leq \mu \leq 65.08$ is either correct with probability 1 or incorrect with probability 1. By and large, the correct interpretation of a $100(1-\alpha)\%$ confidence interval for a population parameter is that:

if a very large number of random samples are collected and a $100(1-\alpha)\%$ *confidence interval for the population parameter is computed from each sample, then* $100(1-\alpha)$ *percent of these intervals will contain the true value of the population parameter with confidence* $100(1-\alpha)$.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

In terms that are loose and respecting the frequency approach to probability, for our case, this general interpretation says

we don't know if the statement $63.84 \leq \mu \leq 65.08$ *is true for this specific sample, but that, in repeated sampling, the method used to obtain the confidence interval for* μ *yields correct statements* 95% *of the times with a 0.95 confidence.*

To illustrate this interpretation from the other end, let us consider a thousand samples taken, and their $100(1-\alpha)\%$ confidence intervals for a specified population parameter constructed. Conceivably, at least fifty of these confidence intervals will fail to contain the true value of the population parameter. Suppose that these were 95% confidence intervals. Then, only 5% of the confidence intervals would fail to contain the true value of the population parameter. This says that 950 of the times would contain the true population parameter of interest with a 0.95 confidence.

We point out that in practice we obtain only one random sample and calculate only one confidence interval. From the preceding standpoint, this confidence interval either contains or does not contain the true population parameter. In the end, one MUST therefore reject the obvious temptation!

Lecture Exercise
For the above example, separately construct a 90% and 99% confidence interval for the population mean.

Further, the length or width of a confidence interval is given by $\ell_2 - \ell_1$. Now, how long are the resulting confidence intervals? Of the three confidence intervals constructed so far, which one is the most precise?

Task: Starting from the cases considered above, what is the general relationship between confidence levels and their precision?

Downloaded by Khann (moin24126@gmail.com)

Remark:
The precision of a confidence interval is inversely proportional to the confidence level. It is desirable to obtain a confidence interval that is short enough for purposes of decision making and that also has adequate confidence. This is easily the reason why the 95% confidence level is the default confidence level chosen by researchers and practitioners.

7.4. One-Sided Confidence Intervals for the Population Mean

Using similar assumptions, one-sided confidence limits for the population mean, μ, are obtained by setting either ℓ_1 to -∞ or ℓ_2 to ∞ and replacing $z_{\frac{\alpha}{2}}$ by z_α.

We therefore have a $100(1-\alpha)\%$ upper confidence limit for μ given by

$$\mu \leq \overline{x} + z_\alpha \times \frac{\sigma}{\sqrt{n}},$$

and a $100(1-\alpha)\%$ lower confidence limit for μ given by

$$\overline{x} - z_\alpha \times \frac{\sigma}{\sqrt{n}} \leq \mu.$$

Lecture Exercise
For the data in the above example, construct the $90\%, 95\%, 99\%$ lower-, and upper- confidence limits. What observations can you make?

Case 2: *Unknown Population Variance*

(a): **Large Samples** ($n > 30$)

It was assumed in the foregoing discussion that the population distribution is normal with an unknown μ and a known standard deviation σ. However, these assumptions may be dropped when dealing with *large-samples*.

Let the observations $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{n}$ be a random sample from a population with unknown mean, μ and an unknown variance, σ^2 . If *n* is large, then

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

and it follows that

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0, 1).$$

In this case n is large and so it is permissible to replace the unknown σ by s. This has close to no effect on the distribution of Z.

For large n, the quantity

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s}$$

follows a standard normal distribution with mean zero and unit standard deviation.

Then the $100(1-\alpha)\%$ confidence interval for μ is

$$\overline{x} - z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}} \leq \ \mu \leq \overline{x} + z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$$

which is true regardless of the sample's underlying distribution.

Example

A study was carried out in Zimbabwe to investigate pollutant contamination in small fish. A sample of small fish was selected from 53 rivers across the country and the pollutant concentration in the muscle tissue was measured (ppm). The pollutant concentration values are shown below. Construct a 95% confidence interval for the population mean, μ.

1.230	1.330	0.040	0.044	1.200	0.270	0.490	0.190	0.940	0.520	0.830
0.810	0.710	0.500	0.490	1.160	0.050	0.150	0.400	0.190	0.650	0.770
1.080	0.980	0.630	0.560	0.410	0.730	0.430	0.590	0.340	0.340	0.270
0.840	0.500	0.340	0.280	0.340	0.250	0.750	0.870	0.560	0.100	0.170
0.180	0.190	0.040	0.490	0.270	1.100	0.160	0.210	0.860		

Solution

Since $n > 30$, then the 95% confidence interval for μ is

$$0.5250 - 1.96 \times \frac{0.3486}{\sqrt{53}} \leq \mu \leq 0.5250 + 1.96 \times \frac{0.3486}{\sqrt{53}}$$

which simplifies to

$$0.431 \leq \mu \leq 0.619$$

Lecture Exercise

Downloaded by Khann (moin24126@gmail.com)

Construct the 90% and the 99% CI for μ using the above data. Further, using the above data construct the 90%, 95%, and the 99% lower- and upper- CI for the population mean.

(b): **Small Samples** ($n \leq 30$)

It is now necessary to introduce a new confidence interval construction procedure that addresses the scenario of *small samples*. In many cases, it is reasonable to assume that the underlying distribution is normal and that moderate departure from normality will have little effect on validity of the result.

Remark: In the equally likely event that the assumption is unreasonable, an alternative (*not discussed here*) is to use the non-parametric procedures which are valid regardless of underlying populations.

For our purposes, it will be reasonable to assume that the population of interest is normal with an unknown mean, μ, and an unknown variance, σ^2. A small random sample of size n is drawn. Let $\overline{X}$ and S^2 be the sample mean and sample variance, respectively. We wish to construct a two-sided confidence interval on μ . The population variance, σ^2, is unknown and it is a reasonable procedure to use S^2 to estimate σ^2 . Then the random variable Z is replaced with T which is given by

$$T = \frac{\sqrt{n}(\overline{X} - \mu)}{s}$$

which is a random variable that follows the student's t distribution with $\mathrm{n} - 1$ degrees of freedom which are associated with the estimated standard deviation.

Notation
We let $\mathrm{t}_{\alpha,\mathrm{n}-1}$ and $\mathrm{t}_{\frac{\alpha}{2},\mathrm{n}-1}$ be the value of the random variable T with $\mathrm{n} - 1$ degrees of freedom above which we find a probability α or $\frac{\alpha}{2}$ respectively.

The $100(1-\alpha)\%$ CI for μ is given by

$$\overline{x} - t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}} \leq \ \mu \leq \overline{x} + t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

where $t_{\frac{\alpha}{2},n-1}$ is the upper $100\frac{\alpha}{2}$ percentage point of the *t*- distribution with $\mathrm{n}-1$ degrees of freedom.

Example
Consider the following data obtained from a local Transport Logistics company.

19.8	10.1	14.9	7.5	15.4	15.4	15.4	18.5	7.9	12.7	11.9
11.4	11.4	14.1	17.6	16.7	15.8	19.5	8.8	13.6	11.9	11.4

Construct a 95% confidence interval for the population mean, μ.

Solution
Since our sample is small, $n = 22$, then the 95% confidence interval for the population mean is given by

$$\overline{x} - t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

Substituting yields

$$13.71 - 2.080 \times \frac{3.55}{\sqrt{22}} \leq \mu \leq 13.71 + 2.080 \times \frac{3.55}{\sqrt{22}}$$

and simplifying we have

$$12.1 \leq \mu \leq 15.3$$

as the 95% confidence interval for μ.

Lecture Exercise
For the above data,construct the 90% and the 99% confidence intervals on the population mean and interpret the two confidence intervals. Further, construct the 90%, the 95% and the 99% lower - and upper - confidence limits. Give an interpretation of each and all of them.

Remark: One-sided confidence intervals for the mean of a normal population are constructed by choosing the appropriate lower- or upper-confidence limit and then replacing $t_{\frac{\alpha}{2},n-1}$ by $t_{\alpha,n-1}$.

Downloaded by Khann (moin24126@gmail.com)

7.5. Confidence Interval for the Population Proportion

Suppose that a random sample of size *n*, large *n*, has been taken from a large population and that *x* but less than *n* observations in this sample belong to a class of interest. Then $\widehat{p}$ calculated as $\frac{x}{n}$ is a point estimator of the proportion of the population *p* that belongs to this class. It is noted that *n* and *p* are the parameters of a binomial distribution (*refer to earlier discussions*). The sampling distribution of $\widehat{p}$ is approximately normal with mean *p* and variance $\frac{p(1-p)}{n}$ if *p* is not too close to either 0 or 1 and if *n* is relatively large. To apply this, it is required that *np* and *n*(1-*p*) be greater than or equal to 5. We are saying that: If *n* is large, then the distribution of

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1).$$

For large samples, which usually is the case when dealing with proportions, a satisfactory $100(1-\alpha)\%$ confidence interval on the population proportion *p* is

$$\widehat{p} - z_{\frac{\alpha}{2}} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq \ p \leq \widehat{p} + z_{\frac{\alpha}{2}} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

where $\widehat{p}$ is the point estimate of p, and $z_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ probability point of the standard normal distribution.

Example
In a random sample of 85 stone sculptures, 10 have a surface finish that is rougher than the expected. Construct a 95% confidence interval for the population proportion of stone sculptures with a surface finish that is rougher than the expected.

Solution
A 95% two-sided confidence interval for *p* is

$$0.12 - 1.96 \times \sqrt{\frac{0.12(1-0.12)}{85}} \leq \ p \leq 0.12 + 1.96 \times \sqrt{\frac{0.12(1-0.12)}{85}}$$

which simplifies to

$$0.05 \leq \ p \leq 0.19$$

Remark: The one-sided lower - and upper - confidence intervals are respec-

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

tively given as

$$\widehat{p} - z_\alpha \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq p$$

and

$$p \leq \widehat{p} + z_\alpha \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

Lecture Exercise
In the above example, construct and interpret the 95% and the 99% lower - and upper - confidence limits for the population proportion.

7.6. Confidence Interval for the Population Variance

Let $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_\mathrm{n}$ be a random sample from a normal distribution with mean μ and variance σ^2, and let S^2 be the sample variance. Then the random variable

$$V = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-square (χ^2) distribution with $\mathrm{n}-1$ degrees of freedom.

Now, if s^2 is the sample variance from a random sample of *n* observations from a normal distribution with unknown variance, σ^2, then a $100(1-\alpha)\%$ confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

where $\chi^2_{\frac{\alpha}{2},n-1}$ and $\chi^2_{1-\frac{\alpha}{2},n-1}$ are the upper and lower $100\frac{\alpha}{2}$ percentage points of the χ^2 distribution with $\mathrm{n}-1$ degrees of freedom, respectively.

Illustration
An Entrepreneur has got an automatic filling machine that she uses to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of s^2 equal to 0.0153 $(fluid\ ounces)^2$. Assume that the fill volume is normally distributed. Then a 95% upper- CI is

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}$$

substituting yields

$$\sigma^2 \leq \frac{(20-1)\times 0.0153}{\chi^2_{1-0.05,20-1}}$$

Downloaded by Khann (moin24126@gmail.com)

simplifying we have

$$\sigma^2 \leq \frac{19 \times 0.0153}{\chi^2_{0.95,19}}$$

where $\chi^2_{0.95,19}$ is 10.117.

so we get

$$\sigma^2 \leq \frac{19 \times 0.0153}{10.117}$$

giving

$$\sigma^2 \leq 0.0287$$

NB: The statistical tables round off 10.117 to 3 s.f.

7.7. Confidence Interval for the Population Standard Deviation

The one-sided lower and upper- confidence intervals for σ^2 are

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \leq \sigma^2$$

and

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}$$

Remark: Clearly, the lower- and upper- confidence intervals/ limits for σ are the square roots of the corresponding limits in the above equations.

We state that $\sigma^2 \leq 0.0287,$ is converted into an upper - confidence interval/ limit for the population standard deviation σ by taking the square root of both sides. The resulting 95% confidence interval is $\sigma \leq 0.17.$

Lecture Exercise
Using the information from the above illustration, construct a 90% lower- and upper- confidence limits for the population standard deviation, σ.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

7.8. Confidence Interval for the Difference of Two Populations Means

The overall assumption remains in place. And, the same with everything else. We are simply considering two populations and constructing confidence intervals for the difference in two population means, $\mu_1 - \mu_2$.

7.8.1. Case 1: Known Population Variance

Illustrative Example
An entrepreneur is interested in reducing the drying time of a wall paint. Two formulations of the paint are tested; formulation 1 is the standard, and formulation 2 has a new drying ingredient that should reduce the drying time. From experience, it is known that the standard deviation of drying time is 8 minutes, and this inherent variability should be unaffected by the addition of the new ingredient. Ten specimens are painted with formulation 1, and another 10 specimens are painted with formulation 2; the 20 specimens are painted in random order. The two sample mean drying times are 121 minutes and 112 minutes, respectively. Construct a 99% confidence interval for the difference in the two population means.

Solution
To be provided in the lecture.

7.8.2. Case 2: Unknown (but assumed Equal) Population Variances

This is called the **Homogeneous Variance Assumption**.

Illustrative Example
The following data is from two populations, A and B. Ten samples from A had a mean of 90.0 with a sample standard deviation of $s_1 = 5.0$, while 15 samples from B had a mean of 87.0 with a sample standard deviation of $s_2 = 4.0$. Assume that the populations, A and B are normally distributed and that both normal populations have the same standard deviation. Construct a 95% confidence interval on the difference in the two population means.

Downloaded by Khann (moin24126@gmail.com)

Solution
To be provided in the lecture.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 8

Hypothesis Testing

8.1. Important Definitions, and Critical Clarifications

Hypotheses
A hypothesis is a statement about a population. Testing of hypotheses evaluates two hypotheses called the null and the alternative denoted H_0 and H_1 respectively. An H_0 is the assertion that a population parameter takes on a particular value. On the other hand, an H_1 expresses the way in which the value of a population parameter may deviate from that specified under H_0. The direction of deviation may be specified (one - sided/ tailed tests) or may not be specified(two - sided/ tailed tests).

We take time to point out that the language and grammar of testing of hypotheses does not use the word *"accept"* or any of its numerous synonyms. This is beyond semantics. To say one *"accepts"* the null hypothesis is to imply that they have proved the null hypothesis to be true. This practice is incorrect. The null hypothesis is the claim that is usually set up with the expectation of rejecting it. The null hypothesis is assumed true until proven otherwise. If the weight of evidence points to the belief that the null hypothesis is unlikely with high probability, then there exists a statistical basis upon which we may reject the null hypothesis. The design of hypotheses tests is such that they are with the null hypothesis until there is enough evidence that suggest support for the alternative hypothesis. Clearly, the design is never about selecting the more likely of the two hypotheses. Let's take this to our legal system. One is considered not guilty until proven otherwise. It is the job of the prosecutor to build a case i.e. put evidence before the court of law that the person in question is guilty. The jury or the judge will give their verdict as guilty or not guilty but will **NEVER** give their verdict with an import of being *innocent*. By and large, the courts of law are a classical example of constant testing of hypotheses procedure. So, let it be clear that on the basis of the data from the sample, we

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

either *reject the null hypothesis* or *fail to reject the null hypothesis*.

In the words of R. A. Fisher:
In relation to any experiment we may speak of ... the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Remark 1: The H_0 reflects the position of no change and will always be worded as an equality.

Remark 2: The language which implies "acceptance" of the null hypothesis is both misleading and against the grammar of the testing of hypotheses.

Test Statistic
This is a value calculated from sample data and is used to decide on rejecting H_0.

Critical Region
This is a range of values which is such that when the test statistic falls into it then H_0 would be rejected.

Critical Value
Is a value that separates the rejection region and the non-rejection region.

Type I error
Occurs when a true null hypothesis is rejected. A null hypothesis is rejected when in actual fact it is true.

Type II error
It occurs when a false null hypothesis is not rejected. Alternatively, it is when a null hypothesis is not rejected when in actual fact it is false.

Level of significance of a Test
Is the probability of making a type I error expressed as a percentage. It is denoted by α.

Downloaded by Khann (moin24126@gmail.com)

Power of a Statistical Test
It is the probability that the testing of hypotheses procedure rejects the null hypothesis when the null hypotheses is indeed false.

8.2. General Procedure on Hypotheses Testing

The following steps are recommended in applying the testing of hypotheses procedure.

- From the problem context, identify the parameter of interest.
- Clearly state the hypotheses i.e. H_0 and H_1.
- Identify or choose the level of significance, α.
- Determine an appropriate test statistic.
- Obtain the critical value from appropriate tables.
- Compute the test statistic by substituting necessary *statistics* into an appropriate equation.
- Decide on the basis of a decision criterion that rejects H_0 if, upon comparison, the test statistic is more extreme than a critical value.
- Conclude on the basis of the decision's import, and report in the context of the problem.

8.3. Hypothesis Testing Concerning the Population Mean

8.3.1. Case 1: Known Population Variance

The overall normality assumption is made. In this case, we further assume that the population variance, σ^2, is known. The sample mean $\overline{X}$ which is a point estimator of μ is a random variable with population mean μ and population variance $\frac{\sigma^2}{n}$. The test statistic is

$$Z_{cal} = \frac{\sqrt{n}(\overline{x} - \mu_0)}{\sigma} \sim N(0, 1)$$

Exercise
Consider the following data where the population mean is claimed to be 50. σ = 2, α = 0.05, $n = 25$, and $\overline{x} = 51.3$ What conclusions should be drawn about the

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

claim?

8.3.2. Guidelines to the Expected Solution

It is given that the population is normal and the population standard deviation is known. $Z-$ score is the test statistic. The testing of hypothesis procedure is two sided. At the 0.05 level of significance and based on the sample evidence, we conclude that the population mean is different from 50.

Exercise
For the above exercise, instead of using the testing of hypothesis procedure, construct a 95% confidence interval. Test the same hypothesis using the confidence interval. Is the value specified under H_0 contained in the confidence interval? Or, is zero contained in the confidence interval? What conclusions should be drawn?

Hint on the Decision Criterion
If the value specified under H_0 is contained in the confidence interval, then we fail to reject H_0. And, if zero is contained in the confidence interval, then we fail to reject H_0.

8.3.3. Case 2: Unknown Population Variance

(a) **Large Samples Scenario** ($n > 30$)
The Z- score. We find the critical value using the level of significance specified or 0.05 if not specified on the basis of the test being a single-sided or a two sided.

Example
Let the mean cost of an Introduction to Statistics textbook be μ. In testing the claim that the population mean is not USD34.50 a sample of 36 current textbooks had selling costs with a sample mean USD32.00 and a sample standard deviation of USD6.30. Using a 10% level of significance, what conclusion can be made?

Downloaded by Khann (moin24126@gmail.com)

Solution
A two-tailed test, $n > 30$, $\alpha = 0.1$ and, thus, the critical value is ± 1.96. Detailed solution in the lecture.

(b) **Small Sample Scenario** (*$n \leq 30$*)
We consider now the case of hypothesis testing on the mean of a population with an unknown variance, σ^2. The test statistic is

$$t_{cal} = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}$$

which follows a *t* - distribution with $\mathrm{n} - 1$ degrees of freedom.

Exercise
The increased availability of light materials with high strength has revolutionized the design and manufacture of golf clubs, particularly drivers. Clubs with hollow heads and very thin faces can result in much longer tee shots, especially for players of modest skills. This is due partly to the spring-like effect that the thin face imparts to the ball. Firing a golf ball at the head of the club and measuring the ratio of the outgoing velocity of the ball to the incoming velocity can quantify this spring-like effect. The ratio of velocities is called the coefficient of restitution of the club. An experiment was performed in which 15 drivers produced by a particular club maker were selected at random and their coefficients of restitution measured. In the experiment the golf balls were fired from an air cannon so that the incoming velocity and spin rate of the ball could be precisely controlled. The sample mean and sample standard deviation are $\overline{x}$ = 0.83725 and s = 0.02456. Determine if there is evidence at the α = 0.05 level to support the claim that the mean coefficient of restitution exceeds 0.82.

Guidelines to the Expected Solution
The mean is the parameter of interest. The population standard deviation is unknown and the sample size is small. Therefore, the appropriate test statistic to be used is the *t* - statistic and the corresponding critical value is $\mathrm{t_{crit}}$ = 1.76.

Exercise
For the above exercise, instead of using the testing of hypothesis procedure, construct a 95% confidence interval. Test the same hypothesis using the confi-

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

dence interval approach.

8.4. Hypothesis Testing concerning the Population Proportion

Illustrative Example
The advertised claim for batteries for cell phones is set at 48 operating hours, with proper charging procedures. A study of 5000 batteries is carried out and 15 stop operating prior to 48 hours. Do these experimental results support the claim that less than 0.2 percent of the company's batteries will fail during the advertised time period, with proper charging procedures? Use a hypothesis testing procedure with α = 0.01. Is the conclusion the same at the 10% level of significance?

Solution
We are testing H_0: p = 0.002 against H_1: p < 0.002 with $\widehat{p} = \frac{15}{5000} = 0.003$.

NB: By claiming a value of the population proportion to be 0.002 implies that the population proportion. Hence letting p_0 to be the hypothesised population proportion value yields

$$Z_{cal} = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 1.5827223454 < Z_{crit} = Z_{0.01} = 2.3263$$

We **fail to reject** H_0 and conclude that, at the 1% level of significance, there is no enough evidence to suggest that less than 0.2 percent of the company's batteries will fail during the advertised time period, with proper charging procedures.

Exercise
Let p be the proportion of new car loans having a 48 months period. In some year p = 0.74. Suppose it is believed that this has declined and accordingly we wish to test this belief using a 1% level of significance. What is the conclusion if 350 of a sample of 500 new car loans have a time period of 48 months?

Downloaded by Khann (moin24126@gmail.com)

8.5. Comparison of Two Populations

We now extend the previous one population results to the difference of means for two populations.

8.5.1. Hypothesis Testing concerning the Difference of Two Population Means

Case 1: *Known Population Variances*

Example Consider the following gasoline mileages of two makes of light trucks. The trucks 1 and 2 have the population means and populations standard deviations 28 and 6, and 24 and 9 respectively. If 35 of truck 1 and 40 of truck 2 are tested, test the claim that the mean difference is 4.

Solution
Exercise in the lecture.

Remark: In inferential applications the population variances σ_1^2 and σ_2^2 are generally not known and must be estimated by s_1^2 and s_2^2. The standard error is estimated by

$$s.e. = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Case 2: *Unknown Population Variance, and Small Sample* ($n_1 + n_2 \leq 31$)

We assume that the variances of both distributions σ_1^2 and σ_2^2 are unknown but equal. This common variance is estimated by a quantity called *pooled variance* denoted s_p^2 and calculated as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Thus, the test statistic is given by

$$t_{cal} = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{s_p^2[\frac{1}{n_1} + \frac{1}{n_2}]}}$$

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

which follows a *t*- distribution with $\mathrm{n}_1 + \mathrm{n}_2 - 2$ degrees of freedom.

Example
Consider the following data. $n_1 = 10$, $\overline{x_1} = 90$, $s_1 = 5$, $n_2 = 15$, $\overline{x_2} = 87$ and $s_2 = 4$. Assume that the populations are normally distributed and that both populations have the same standard deviation. At the 5% level of significance, can we conclude that there is a difference in the two population means?

Solution
Left as an exercise for the lecture.

8.6. Independent Samples and Dependent/ Paired Samples

In testing for the equality of two population means, we may choose to select two random samples one from each population and compare their means. If these sample means exhibit a difference, then we reject the null hypothesis that $H_0 : \mu_1 - \mu_2 = 0$. Another approach is to try and match the subjects from the two populations according to variables which will be expected to have an influence on the variable under study. The two samples are no longer independent and the inferences are now based on the differences of the observations from the matched pairs.

Case 3: *Independent Samples* An illustrative example will be vital in exposing the testing of hypothesis procedure.

Example
Samples of two brands of pork sausage are tested for their fat content. The results of the percentage of fat are summarised as follows: Brand A (n=50, $\overline{x}$=26.0, s=9.0) and Brand B (n=46, $\overline{x}$=29.3, s=8.0). Can we conclude that there is sufficient evidence to suggest that there is a difference in the fat contented of the two brands of pork sausage? Use a 5% level of significance.

Downloaded by Khann (moin24126@gmail.com)

Solution
Left as an exercise for the lecture.

Case 4: *Dependent / Paired Samples*
Given two paired samples $\mathrm{X}_{11}, \mathrm{X}_{12}, \cdots, \mathrm{X}_{1\mathrm{n}_1}$ and $\mathrm{X}_{21}, \mathrm{X}_{22}, \cdots, \mathrm{X}_{2\mathrm{n}_2}$ we form a single sample of the differences $\mathrm{d}_1, \mathrm{d}_2, \cdots, \mathrm{d}_\mathrm{n}$ where $\mathrm{d}_1 = \mathrm{X}_{11} - \mathrm{X}_{21}, \mathrm{d}_2 = \mathrm{X}_{12} - \mathrm{X}_{22}, \cdots, \mathrm{d}_\mathrm{n} = \mathrm{X}_{1\mathrm{n}_1} - \mathrm{X}$

For the new single sample, we find its mean, $\overline{d}$, that estimates the population mean for the differences, μ_d and standard deviation, $\mathrm{s_d}$. Assuming that the original populations are normally distributed with equal means i.e. $\mu_1 = \mu_2$ and equal variances, the population mean for the differences μ_d is zero and a standard error that is estimated by $\frac{s_d}{\sqrt{n}}$.

The test statistic in this case is $t_{cal} = \frac{\overline{d}}{s.e.}$

The hypotheses tests concerning μ_1 and μ_2 are now based on the sample mean using the single sample and we have a modified null hypothesis, H_0: $\mu_d = 0$ against an appropriate alternative hypothesis as instructed by the situation.

Example
Five automachines are tested for wind resistance with two types of grills. Their drag coefficients were determined and recorded as follows.

Automachine	1	2	3	4	5
Grill A	0.47	0.46	0.40	0.44	0.43
Grill B	0.50	0.45	0.47	0.44	0.48

Using a 5% level of significance test for the difference in the drag coefficients due to type of grill.

Solution
Left as an exercise during the lecture.

8.6.1. Advantages of Paired Comparisons

- By pairing, we remove the additional source of variation (i.e. the difference between experimental units) and hence reduce the random variation

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

as measured by $\mathrm{s}_{\mathrm{p}}^2$ in the case without pairing and $\mathrm{s}_{\mathrm{d}}^2$ in the case of pairing. Therefore, $\mathrm{s}_{\mathrm{d}}^2 > \mathrm{s}_{\mathrm{p}}^2$ implies a gain in precision due to pairing.

- The confidence interval based on the paired comparison is much narrower than that from two sample analysis using unpaired observations. This also implies a gain in precision due to pairing.

- It may be less expensive since in most cases fewer experimental units are used when compared to a two sample design.

8.6.2. Disadvantages of Paired Comparisons

- There is a substantial loss in the degrees of freedom in a paired comparison than in a two sample *t* - test.

- A rest period may be required between applying the first and second treatment in order to minimise the carry over effect from the first treatment. Even, then the carry over effect may not be completely eliminated.

8.7. Test Procedure concerning the Difference of two Population Proportions

Suppose that two independent random samples of sizes n_1 and n_2 are taken from two populations, and let x_1 and x_2 represent the number of observations that belong to the class of interest in sample 1 and sample 2, respectively. In testing the hypotheses

$$H_0 : p_1 - p_2 = 0$$
$$H_1 : p_1 - p_2 \neq 0,$$

the test statistic is

$$Z_{cal} = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

which is approximately standard normal.

If $H_0 : p_1 - p_2 = 0$ is true, then $p_1 = p_2$. Thus $p_1 = p_2 = p$ such that the test statistic becomes

$$Z_{cal} = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{p(1-p)[\frac{1}{n_1} + \frac{1}{n_2}]}}$$

Downloaded by Khann (moin24126@gmail.com)

which still is approximately standard normal. The common population proportion, $p,$ is estimated by

$$\widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

Assuming that $H_0 : p_1 - p_2 = 0$ is true, the test statistic is therefore given by

$$Z_{cal} = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\widehat{p}(1-\widehat{p})[\frac{1}{n_1} + \frac{1}{n_2}]}}$$

Example
Consider the following situation in which comparison is made of two concept exposition methods. Method A is the standard and method B is the proposed. A class of 200 CUMT105 students at the Chinhoyi University of Technology is used. The students were randomly assigned to two groups of equal size. One group was exposed to method A, and the other group was exposed to method B. At the end of the semester, 19 of the students exposed to method B showed improvement, while 27 of those exposed to method A improved. At the 5% level of significance, is there sufficient reason to believe that method A is effective in concept exposition?

Solution
First, we state the hypotheses:

$$H_0 : p_A - p_B = 0$$
$$H_1 : p_A - p_B \neq 0$$

Then, we extract the given data: $n_A = n_B = 100$, $\widehat{p_A} = 0.27$, $\widehat{p_B} = 0.19$, $x_A = 27$, and $x_B = 19$. Thus, $\widehat{p} = 0.23$.

The test statistic and the critical value are $Z_{cal} = 1.35$ and $Z_{crit} = 1.96$ respectively.

After comparing Z_{cal} and Z_{crit}, the decision is that we **fail to reject** H_0. From this decision, we therefore conclude that, at the 5% level of significance, there is no sufficient evidence to support the assertion that method A is effective in concept exposition.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Exercise
A study is made of business support of the immigration enforcement practices. Suppose 73% of a sample of 300 cross border traders and 64% of the light manufacturers said they fully supported the policies being proposed. Is there sufficient evidence to conclude that the proposed policies are equally supported by the two groups sampled. Use a 1% level of significance.

8.8. Tests for Independence

Tests for independence are performed on categorical data such as when testing for independence of opinion on a public policy and gender. The data is contained in what is called a contingency table. The hypotheses are tested using a Chi - square test statistic, χ^2_{cal}.

Illustrative example
A company operates four machines three shifts each day. From production records, the following data on the number of breakdowns are collected:

	Machines			
Shifts	A	B	C	D
1	4	3	2	1
2	3	1	9	4
3	1	1	6	0

Using 5% level of significance, test the hypothesis that breakdowns are independent of the shift.

Solution: To be provided in the lecture.

Exercise
Grades in Statistics and Communication Skills taken simultaneously were recorded as follows for a particular group of students.

	Com. Skills Grade			
Stats Grade	1	2.1	2.2	Other
1	25	6	17	13
2.1	17	16	15	6
2.2	18	4	18	10
Other	10	8	11	20

Are the grades in Statistics and Communication Skills related? Use α = 0.01

Downloaded by Khann (moin24126@gmail.com)

8.9. Ending Remark(s)

It is required that one may demonstrate that hypothesis testing and confidence intervals are equivalent procedures insofar as decision making or inference about population parameters is concerned. However, each procedure presents different insights. What is the major difference between these two *cousin* procedures?

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 9

Regression Analysis

9.1. Introduction

It is important to note that the approach used here first exposes the useful concepts of the regression analysis technique, gives an illustrative example on the application of these concepts, and then wraps up with a practice question.

Many problems that are encountered in everyday life involve exploring the relationships between two or more variables. Without attempting to formally define what regression analysis is, regression analysis is a statistical tool that is very useful for these types of problems. For example, in the clothing industry, the sales obtained from selling particular designer outfits is related to the amount of time spent advertising the label. Regression analysis can be used to build a model to predict the sales given the amount of time devoted to advertising the label. In the sciences, regression analysis models can be used for process optimization. For instance, finding the temperature levels that maximise yield, or for puproses of process control.

After non-superficial, but serious and rigorous studying of this chapter, the student is expected to be able to use simple linear regression for building models to everyday data, apply the method of least squares to estimate the parameters in a linear regression model, use the fitted regression model to make a prediction of a future observation and interpret the scatter plot, the correlation coefficient, the coefficient of determination, and the regression parameters.

9.2. Uses of Regression Analysis

The uses of regression are but not limited to:

- understanding underlying processes

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

- prediction
- forecasting
- optimisation
- control purposes

9.3. Abuses of Regression Analysis

Regression analysis is widely used and frequently misused. Several common abuses of regression include developing statistically significant relationships among variables that are completely unrelated in a *cause - effect* sense. A strong observed association between variables does not necessarily translate into a cause - effect relationship between the variables. Therefore, care must be exercised when choosing variables on which to perform regression analysis.

Regression relationships are valid only for values of the explanatory variable within the range of the original data. The linear relationship that we have assumed may be valid over the original range of X, but may unlikely remain so as we extrapolate i.e. if we use values of X beyond the range in question to estimate the value of Y. Alternatively put, as we stride from the range of the values of X for which data were collected, our certainty about the validity of the assumed model tend to fade away. We caution that linear regression models are not necessarily valid for extrapolation purposes. Clearly, this is not saying *NO to extrapolation*. Note that in many life situations extrapolation of a regression model may be the only way to approach a given problem. We are strongly warning that there is need to be alive to the potential abuses of the treasure. To dilute the preceding a bit, a modest extrapolation may be quite fine in most situations, however large extrapolations will almost always produce unacceptable results.

Well, we will concentrate on two random variables: the explanatory variable or the independent variable or the cause variable among a host of other names which is denoted X, and the response variable or the dependent variable or the effect variable among a host of other names denoted by Y. These two variables vary together. X causes Y to vary or X explains the response in Y. Such situations are modeled using the *simple linear regression analysis* technique be-

Downloaded by Khann (moin24126@gmail.com)

cause they have only one explanatory variable or independent variable. Specifically, this will be our focus in this chapter.

9.4. The Simple Linear Regression Model

The simple linear regression model is an equation of a straight line given by

$$Y = a + bX + \epsilon$$

where Y is the response / dependent variable, a is a regression coefficient / regression parameter called the *intercept*, b is a regression coefficient / regression parameter called the *slope*, X is the explanatory/ independent variable and e is the random error term.

The random error term follows a normal distribution with a mean zero and an unknown variance σ^2. For completeness, we state that the random errors corresponding to different observations are also assumed to be uncorrelated random variables. To determine the appropriateness of employing simple linear regression we use (1) the scatter plot, and or (2) the correlation coefficient techniques.

9.4.1. The Scatter Plot

The choice of the model is based on inspection of a scatter diagram. We merely use our eyes to inspect the nature of the relationship exhibited by the points. The slope of the points will tell us the direction of the relationship and the distances between the points will tell us the nature of the magnitude of the relationship. Invariably, we note that there is a deep seated tendency to join the points on the scatter plot. We therefore take the unprecedented step of telling you what is *not correct*. Let us be clear that the points need not be joined and that there is **NO** line of whatever form to be drawn on a scatter diagram.

9.4.2. The Correlation Coefficient

A correlation coefficient measures the strength and direction of the relationship between two variables that vary together. We will compute the Pearson

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

product moment correlation coefficient denoted by *r* and given by

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{9.1}$$

Note that $-1 \leq r \leq 1$. Though seemingly intimidating, this formula is quite user friendly. Employ it and discover more!

Interpretation of *r*

When interpreting *r* note should be taken to mention the magnitude / size of the correlation and the direction of the linear relationship. The absence of the other component renders the interpretation incomplete. We now give typical interpretations for some values of *r*. If *r* = -1, then we say that there is a perfect negative linear relationship between X and Y. If *r* = 1, then there is a perfect positive linear relationship between X and Y. If *r* = 0, then there is a no linear relationship between X and Y. Note that this implies that other forms of relationships may best model the situation. If *r* = -0.9, then that there is a very strong negative linear relationship between X and Y. If *r* = 0.9, then there is a very strong positive linear relationship between X and Y. If *r* = -0.75, then that there is a strong negative linear relationship between X and Y. If *r* = 0.75, then there is a strong positive linear relationship between X and Y. If *r* = -0.5, then there is a fair negative linear relationship between X and Y. If *r* = 5, then there is a fair positive linear relationship between X and Y. Any value of *r* that is less than -0.5 or 0.5 says that there is a weak negative or weak positive linear relationship between X and Y. Thus, using simple linear regression to model the situation is not advisable.

Remark: The interpretation of *r* must clearly state the magnitude/ size and direction of the relationship between the random variables X and Y.

9.4.3. Regression Equation

Having established that a linear relationship exists between the random variables X and Y, we proceed to fit the linear regression model/ line/ equation. To fit a regression model is to estimate the regression coefficients a and b. The estimates are denoted $\widehat{a}$ and $\widehat{b}$. The fitted model is written in the form

$$\widehat{Y} = \widehat{a} + \widehat{b}X$$

Downloaded by Khann (moin24126@gmail.com)

Now, we have fitted a model and we wish to determine how good it is and then use it for prediction of new values for the system in question. To determine how good our model is we calculate the values of the response variable in reverse for each and every value of the explanatory variable and then note the difference. This difference obtained by subtracting the value of the fitted model from the actually observed is the error in our model for that observation and it is called the *residual*. By performing what is called *residual analysis* we are able to come up with a statement on the adequacy of our fitted regression model.

After establishing the adequacy of our model we then proceed to predict future values of the response variable for the system in question. This is technically called *forecasting*.

Computation of the Regression Coefficients
The method used to compute the regression coefficients is called the *least squares method*.

We first estimate the slope, b as

$$\widehat{b} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

And then we immediately estimate a by

$$\widehat{a} = \overline{y} - \widehat{b}\overline{x}$$

Interpretation of the Regression Coefficients
The intercept, a is the value of Y when X = 0. *The slope,* b indicates the value of Y when X changes by a unit value.

Naturally, how much of the variability in the response variable has been explained by fitting the regression model? To answer this question we need to compute the following coefficient.

9.4.4. Coefficient of Determination, r^2

The coefficient of determination establishes the amount of variability in the response variable that has been explained or accounted for by fitting a regression model. It is obtained by squaring the correlation coefficient, *r* and we have

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

$0 \leq R^2 \leq 1$. Expressing r^2 as a percentage

$$r^2 = r^2 \times 100\%$$

gives the amount of variability in the response variable that has been explained by fitting a regression model.

Illustrative Example

Consider the following set of observations. Take X to be the exploratory variable and Y to be the response variable.

Y	1	0	1	2	5	1	4	6	2	3	5	4	6	8	4
X	60	63	65	70	70	70	80	90	80	80	85	89	90	90	90

1. Draw a scatter plot for the above data. Comment on the suitability of using simple linear regression to describe the relationship.
2. Calculate and comment on the Pearson correlation coefficient.
3. Fit the regression model using the method of least squares. Interpret the regression coefficients.
4. State how much of the variation in Y has been accounted for by fitting the linear regression model.
5. Using the fitted regression model, what is the value of Y when X = 60? What is the residual?
6. What is the value of Y when X = 95?

Solution

A scatter diagram of the above data is shown in the figure below.

Exercise

Consider the following quantities for two random variables X and Y. Let X be the cause variable and Y be the effect variable.

$n = 20, \sum x = 24, \sum y = 1843, \sum y^2 = 170045, \sum x^2 = 29$ and $\sum xy = 2215$

1. Is it appropriate to employ simple linear regression analysis on these data?
2. Fit the regression model using the method of least squares. What is the meaning of the regression coefficients?

Downloaded by Khann (moin24126@gmail.com)

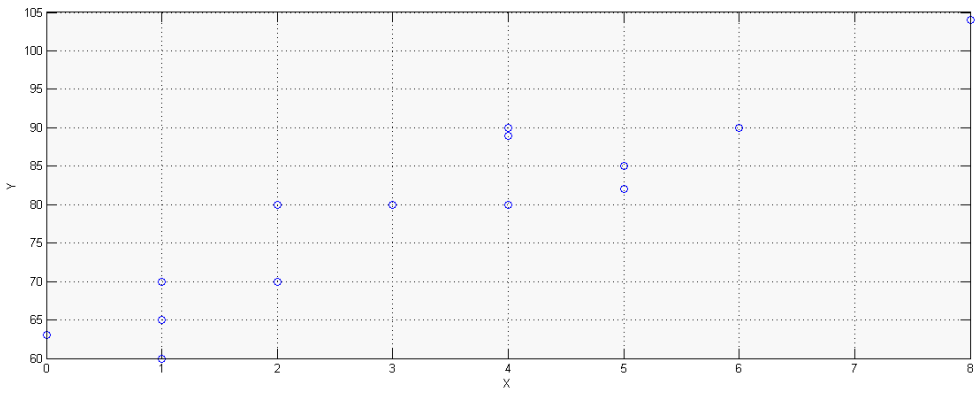

Figure 9.1: A scatter diagram of X and Y values

3. How much of the variability in Y has been explained by fitting the linear regression model above.

4. Using the fitted regression model, what is the value of Y when X = 2? What is the residual?

5. What is the value of Y when X = 25?

6. Comment on the usefulness of the values in parts (d) and (e) given that for twenty observations $\sum x = 24$. **Hint:** You are expected to reflect on the uses and abuses of the regression analysis technique.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Downloaded by Khann (moin24126@gmail.com)

Chapter 10

Index numbers

10.1. Objectives

After reading this chapter, you will be conversant with

1. The Concept of Index Numbers
2. Uses of Index Number
3. Different Types of Index Numbers
4. Aggregates Method of Constructing Index Numbers

10.2. Introduction

Index numbers are today one of the most widely used statistical indicators. Generally used to indicate the state of the economy, index numbers are aptly called barometers of economic activity. Index numbers are used in comparing production, sales or changes exports or imports over a certain period of time. The role-played by index numbers in Indian trade and industry is impossible to ignore. It is a very well known fact that the wage contracts of workers in our country are tied to the cost of living index numbers.

10.3. What is an Index Number?

By definition, an index number is a statistical measure designed to show changes in a variable or a group of related variables with respect to time, geographic location or other characteristics such as income, profession, etc.

10.3.1. Characteristics of an Index Numbers

1. These are expressed as a percentage: Index number is calculated as a ratio of the current value to a base value and expressed as a percentage.

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

It must be clearly understood that the index number for the base year is always 100. An index number is commonly referred to as an index.

2. Index numbers are specialized averages: An index number is an average with a difference. An index number is used for purposes of comparison in cases where the series being compared could be expressed in different units i.e. a manufactured products index (a part of the whole sale price index) is constructed using items like Dairy Products, Sugar, Edible Oils, Tea and Coffee, etc. These items naturally are expressed in different units like sugar in kg, milk in litres, etc. The index number is obtained as a result of an average of all these items, which are expressed in different units. On the other hand, average is a single figure representing a group expressed in the same units.

3. Index numbers measures changes that are not directly measurable: An index number is used for measuring the magnitude of changes in such phenomenon, which are not capable of direct measurement. Index numbers essentially capture the changes in the group of related variables over a period of time. For example, if the index of industrial production is 215.1 in 1992-93 (base year 1980-81) it means that the industrial production in that year was up by 2.15 times compared to 1980-81. But it does not, however, mean that the net increase in the index reflects an equivalent increase in industrial production in all sectors of the industry. Some sectors might have increased their production more than 2.15 times while other sectors may have increased their production only marginally.

10.3.2. Uses of Index Numbers

Index numbers are used for:

1. Establishes trends - Index numbers when analyzed reveal a general trend of the phenomenon under study. For eg. Index numbers of unemployment of the country not only reflects the trends in the phenomenon but are useful in determining factors leading to unemployment.

2. Helps in policy making - It is widely known that the dearness allowances paid to the employees is linked to the cost of living index, generally the consumer price index. From time to time it is the cost of living index, which forms the basis of many a wages agreement between the employees union and the employer. Thus index numbers guide policy making.

3. Determines purchasing power of the dollar Usually index numbers are used to determine the purchasing power of the dollar. Suppose the con-

Downloaded by Khann (moin24126@gmail.com)

sumers price index for urban non-manual employees increased from $100 in 2004 to $202 in 2006, the real purchasing power of the dollar can be found out as follows:

$$\frac{100}{202} = 0.495$$

It indicates that if dollar was worth $100 in 2004 its purchasing power is $49.5 in 2006.

4. Deflates time series data - Index numbers play a vital role in adjusting the original data to reflect reality. For example, nominal income (income at current prices) can be transformed into real income(reflecting the actual purchasing power) by using income deflators. Similarly, assume that industrial production is represented in value terms as a product of volume of production and price. If the subsequent years industrial production were to be higher by 20% in value, the increase may not be as a result of increase in the volume of production as one would have it but because of increase in the price. The inflation which has caused the increase in the series can be eliminated by the usage of an appropriate price index and thus making the series real.

10.4. Types of Index Numbers

There are three principal types of indices, these are:

- Price index,
- Quantity index
- Value index.

1. Price Index - The most frequently used form of index numbers is the price index. A price index compares charges in price of edible oils. If an attempt is being made to compare the prices of edible oils this year to the prices of edible oils last year, it involves, firstly, a comparison of two price situations over time and secondly, the heterogeneity of the edible oils given the various varieties of oils. By constructing a price index number, we are summarizing the price movements of each type of oil in this group of edible oils into a single number called the price index. The Whole Price Index (WPI) and the Consumer Price Index (CPI) are some of the popularly used price indices.

2. Quantity Index - A quantity index measures the change in quantity from one period to another. If in the above example, instead of the price of

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

edible oils, we are interested in the quantum of production of edible oils in those years, then we are comparing quantities in two different years or over a period of time. It is the quantity index that needs to be constructed here. The popular quantity index used in this country and elsewhere is the index of industrial production (HP). The index of industrial production measures the increase or decrease in the level of industrial production in a given period compared to some base period.

3. Value Index - The value index is a combination index. It combines price and quantity changes to present a more spatial comparison. The value index as such measures changes in net monetary worth. Though the value index enables comparison of value of a commodity in a year to the value of that commodity in a base year, it has limited use. Usually value index is used in sales, inventories, foreign trade, etc. Its limited use is owing to the inability of the value index to distinguish the effects of price and quantity separately. What are the methods of constructing Index Numbers?

10.5. Methods of constructing index numbers

There are two approaches for constructing an index number namely

1. Aggregate method

2. Average of relatives method.

Note: The index constructed in either of these methods could be either an unweighted index (and unweighted index is an index where equal weights are implicitly assigned) or a weighted index (a weighted index is an index where weighted are assigned to the various items constituting the index.)

10.5.1. Aggregate Method

Under the aggregates method of constructing an index number, we could have **unweighted** and **weighted** aggregates index.

Unweighted Aggregates Index
An unweighted aggregates index is calculated by totalling the current year/given years elements and then dividing the result by the sum of the same elements during the base period. To construct a price index, the following mathematical

Downloaded by Khann (moin24126@gmail.com)

formula may be used.

$$Unweighted\ Aggregate\ Price\ Index = \frac{\sum P_1}{\sum P_0} \times 100\% \tag{10.1}$$

Where, $\sum P_1$ = Sum of all elements in the composite for current year and $\sum P_0$ = Sum of all elements in the composite for base year.

10.5.2. Merits and demerits of this method

Merit - This is the simplest method of constructing index numbers.

Demerits - It does not consider the relative importance of the various commodities involved. The unweighted index doesnt reflect the reality since the price changes are not linked to any usage/consumption levels.

Example
Construct an unweighted index for the three commodities taking 2010 as the base year.

Commodities	Prices	
	2010 (P_0)	2012 (P_1)
Oranges (Dozen)	20	28
Milk (Ltr)	5	8
Gas	76	100

The unweighted aggregate price index (UAPI) by:

$$\begin{aligned} UAPI &= \frac{\sum P_1}{\sum P_0} \times 100\% \\ &= \frac{136}{101} \times 100\% \\ &= 134.65\% \end{aligned} \tag{10.2}$$

Above we measured changes in general price levels on the basis of changes in prices of a few items. While the year 1990 was taken as the base year, a comparison has been made between the prices of 1992 and that of the base year 1990.

Interpretation
As evident, the price index was 134.65, which means that the prices rose by 34.65 percent from 1990 to 1992. By no means should this price index be

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

interpreted as a reflection of the price changes of all goods and services, as this calculation is a rough estimate. On inclusion of other items/elements and varying weights in the composite, with 1990 as the base year and 1992 as the current year, there is every possibility that the calculated price index would be different from the price index calculated earlier. This factor can be cited as one of the drawbacks of the simple unweighted index. The unweighted index doesnt reflect the reality since the price changes are not linked to any usage/consumption levels. On the other hand, a weighted index attaches weights according to their significance and hence is preferred to the unweighted index.

To make this clear, let us calculate the price index with the same data provided above but by changing the milk consumption from 1 liter to 100 liters. The following table provides the calculation of the price index.

10.5.3. Weighted Aggregates Index

In a weighted aggregates index, weights are assigned according to their significance and consequently the weighted index improves the accuracy of the general price level estimate based on the calculated index. Generally, the level of consumption is taken as a measure of its importance in computing a weighted aggregates index. There are various methods of assigning weights to an index. The more important ones are:

1. Laspeyres Method
2. Paasche Method
3. Fixed Weight Aggregates Method
4. Fishers Ideal Method and

10.5.4. Laspeyres Method

Laspeyres method uses the quantities consumed during the base period in computing the index number. This method is also the most commonly used method which incidentally requires quantity measures for only one period. Laspeyres index can be calculated using the following formula:

$$Laspeyres\ Price\ Index(LPI) = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100\% \tag{10.3}$$

Where, P_1 = Prices in the current year, P_0 = Prices in the base year, Q_0 = Quantities in the base year.

Downloaded by Khann (moin24126@gmail.com)

In general, Laspeyres price index calculates the changes in the aggregate value of the base years list of goods when valued at current year prices. In other words, Laspeyres index measures the difference between the theoretical cost in a given year and the actual cost in the base year of maintaining a standard of living as in the base year. Also, Laspeyres quantity index can be calculated by using the formula:

$$Laspeyres\ Quantity\ Index(LQI) = \frac{\sum P_0Q_1}{\sum P_0Q_0} \times 100\% \tag{10.4}$$

Where, Q_1 = Quantities in the current year and Q_0, P_0 are as defined earlier. Let us understand this with the help of an example. Calculate the Laspeyres Price and Quantity Indices for the following production data.

Product	Production		Prices				
	Q_0	Q_1	P_0	P_1	P_0Q_0	P_1Q_0	P_0Q_1
	1985	1990	1985	1990			
Rice	46.60	58.00	700	910	32620.00	42406.00	40600.00
Sugar	14.57	17.92	620	950	9033.40	13841.50	11110.40
Salt	69.46	85.10	205	300	14239.30	20838.00	17445.50
Wheat	33.84	40.30	330	470	11167.20	15904.80	13299.00

Solution

Laspeyres price index is:

$$\begin{aligned} Laspeyres\ Price\ Index(LPI) &= \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100\% \\ LPI &= \frac{9299030}{6705990} \times 100\% \\ LPI &= 138.67\% \end{aligned}$$

Laspeyres quantity index is:

$$\begin{aligned} Laspeyres\ Quantity\ Index(LQI) &= \frac{\sum P_0Q_1}{\sum P_0Q_0} \times 100\% \\ &= \frac{82454.90}{67059.90} \times 100\% \\ &= 122.96\% \end{aligned}$$

10.5.5. Merits and demerits of Laspeyres method?

Merits - A laspeyres index is simpler in calculation and can be computed once the current year prices are known, as the weight are base year quantities in a

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

price index. This enables us an easy comparability of one index with another.

Demerits - Laspeyres tends to overestimate the rise in prices or has an upward bias. Let us see how it overestimates the rise in prices or has an upward bias. There is usually a decrease in the consumption of those items for which there has been a considerable price hike and the usage of base year quantities will result in assigning too much weight to prices that have increased the most and the net result is that the numerator of the Laspeyres index will be too large. Similarly, when the prices go down, consumers tend to demand more of those items that have declined the most and hence the usage of base period quantities will result in too low weight to prices that have decreased the most and the net result is that the numerator of the Laspeyres index will again be too large. This is a major disadvantage of the Laspeyres index. However, the Laspeyres index remains most popular for reason of its practicability. In most countries, index numbers are constructed by using Laspeyres formula.

10.5.6. Paasches Method

Paasche index is similar to that of computing a Laspeyres index. The difference is that the Paasche method uses quantity measures for the current period rather than for the base period. The Paasche index can be calculated using the following formula.

$$Paasche\ Price\ Index(PPI) = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100\% \tag{10.5}$$

Where, P_1 = Prices in the current year P_0 = Prices in the base year Q_1 = Quantities in the current year. The Paasche quantity index is given by:

$$Paasche\ Quantity\ Index(PQI) = \frac{\sum P_1Q_1}{\sum P_1Q_0} \times 100\% \tag{10.6}$$

10.5.7. Merits and Demerits of Paasches Index

Merit - Paasches index attaches weights according to their significance.

Demerits - Paasche index is not frequently used in practice when the number of commodities is large. This is because for Paasche index, revised weights or quantities must be computed for each year examined. Such information is ei-

Downloaded by Khann (moin24126@gmail.com)

ther unavailable or hard to gather adding to the data collection expense, which makes the index unpopular. Paasche index tends to underestimate the rise in prices or has a downward bias.

Let us understand the Paasches method with the help of an example. The table below represents the calculation of Paasche index. In general Paasche index reflects the change in the aggregate value of the current years (given periods) list of goods when valued at base period prices. From the table below calculate the Paasche price and quantity indices.

	1992		**1993**					
Commodity	Price	Quantity	Price	Quantity	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	3	18	4	15	54	45	72	60
B	5	6	5	9	30	45	30	45
C	4	20	6	26	80	104	120	156
D	1	14	3	15	14	15	42	45
					178	209	264	306

Solution

Paasche Price Index (PPI) is:

$$\begin{aligned} Paasche\ Price\ Index(PPI) &= \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100\% \\ PPI &= 146.41\% \end{aligned}$$

Paasche Quantity Index (PQI)

$$Paasche\ Quantity\ Index(PQI) = \frac{\sum P_1Q_1}{\sum P_1Q_0} \times 100\% PQI = 115.91\%$$

Paasche price index is 146.41%. Laspeyres Price Index, when calculated is 148.31%

The difference between the purchase index and Laspeyres index reflects the change in consumption patterns of the commodities A, B, C and D used in that table. As the weighted aggregates price index for the set of prices was 148.31% using the Laspeyres method and 146.41 using the Paasche method for the same set, it indicates a trend towards less expensive goods. Generally, Laspeyres and Paasche methods tend to produce opposite extremes in index values computed from the same data. The use of Paasche index requires the continuous use of new quantity weights for each period considered. As opposed to the

This document is available free of charge on studocu

Downloaded by Khann (moin24126@gmail.com)

Laspeyres index, Paasche index generally tends to underestimate the prices or has a downward bias. Because people tend to spend less on goods when their prices are rising, the use of the Paasche which bases on current weighting, produces an index which does not estimate the raise in prices rightly showing a downward bias. Since all prices or all quantities do not move in the same order, the goods which have risen in price more than others at a time when prices in general are rising will tend to have current quantities and they will thus have less weight in the Paasche index.

10.6. Fisher Index

Prof. Irving Fisher has proposed a formula for constructing index numbers, as the geometric mean of the Laspeyres and Paasche indices i.e. Fisher's quantity and price index aare calculated as:

$$Fisher's\ Quantity\ Index\ =\ \sqrt{(Laspeyres\ Quantity\ Index \times Paasche\ Quantity\ Index)}$$

$$Fisher's\ Price\ Index\ =\ \sqrt{(Laspeyres\ Price\ Index \times Paasche\ Price\ Index)}$$

The following advantages can be cited in favor of Fishers Index:

1. Theoretically, geometric mean is considered the best average for the construction of index numbers and Fishers index uses geometric mean.
2. As already noted, Laspeyres index and Paasche index indicate opposing characteristics and Fishers index reduces their respective biases. In fact, Fishers ideal index is free from any bias. This has been amply demonstrated by the time reversal and factor reversal tests.
3. Both the current year and base year prices and quantities are taken into account by this index. The Index is not widely used owing to the practical limitations of collecting data. Fishers Ideal Quantity Index can be found out by the formula.

Downloaded by Khann (moin24126@gmail.com)